USING PATIENT PREFERENCES TO ESTIMATE OPTIMAL TREATMENT STRATEGIES FOR COMPETING OUTCOMES

Emily Lynn Butler

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2016

Approved by:

Michael R. Kosorok

Eric B. Laber

Sonia M. Davis

Lloyd J. Edwards

Donglin Zeng

Stephen R. Cole

ABSTRACT

Emily Lynn Butler: Using Patient Preferences to Estimate Optimal Treatment Strategies for Competing Outcomes
(Under the direction of Michael R. Kosorok)

Treatment decisions should be tailored as close as possible to heterogeneous disease populations because diversity permeates through all levels of patient information. This can include genetic makeup, demographic variables, or individual goals as to what qualifies as a successful outcome; hence, patient treatment plans should account for all of these factors. This area of clinical research, coined precision medicine, focuses on combining a multitude of considerations to make treatment decisions as personalized as possible. In this sphere, the statistical contribution involves methodology that most accurately maps patient information to the set of treatment options. While there has been a plethora of research developing personalized treatment plans, they are centralized around creating optimal strategies for only one outcome. There has been limited work done to estimate personalized treatment plans for patients interested in balancing competing outcomes. This work seeks to fill that gap. One way of balancing competing outcomes is to incorporate the patient's preferences regarding these outcomes in the development of a utility function to be used in the estimation procedure. Since it is not possible to directly observe a patient's preference in standardized numerical form, we solve this using a preference elicitation questionnaire in conjunction with item response theory. We derive a posterior estimate of each patient's latent preference information and use it to define a utility function that represents the patient's inherent trade-offs between the outcomes. The optimal treatment choice is that which provides the largest

expected utility for each patient, conditional on the patient's prognostic information. This estimation technique is extended to the multi-stage treatment scenario which requires sequential decision making. Estimating the latent preference value now involves the patient's contentment with results from previous stages along with the evolution of the patient's preferences. Once the composite outcome is defined, Q-learning is used to determine which treatment elicits the largest expected utility given the patient's prognostic information, while assuming that the optimal treatment will be chosen in the future. Finally, to make the estimation technique more flexible, we propose a non-parametric approach to both estimating the latent preference and defining the utility function via monotone splines.

ACKNOWLEDGMENTS

My advisor, Dr. Michael Kosorok, deserves immense acknowledgment and my eternal gratitude for guiding and supporting me not only through the dissertation process, but throughout my entire time at the University of North Carolina. He has not only been a teacher and a mentor, but an invaluable source of encouragement through frustrating times and an unwavering sense of stability within the department. His patience throughout this process is unmatched and is only surpassed by his immense kindness, both on and off campus. His support of my decision to work remotely impacted my education, progress and success tremendously and my only hope is to make him proud by going out into the world and making substantive contributions to the medical and statistical community.

My profuse appreciation is given to Dr. Eric Laber for his guidance, innovation, foresight, and feedback throughout this process. To say this work would not exist without his diligence and contribution is an understatement. I would also like to thank my committee, Dr. Davis, Dr. Edwards, Dr. Zeng and Dr. Cole, for their time, thoughtful input and advice.

They say it takes a village to be successful, and no one knows this more than I. My husband, Matthew, has been my biggest rock since the day I walked on campus and his devotion has only increased as I transitioned to the dissertation process. This research is dedicated to him and his steadfast comfort, strength, support and relief. Matt, I owe you tremendous gratitude for all of the sacrifices and concessions you have made on my behalf; all of this is because of you. My family, Dale, Jerinah and Glen,

have been my cheerleaders since before I can remember and always believed in me when I didn't believe in myself. This accomplishment is not only mine, but is theirs, and I could not have had a better team in my corner. Finally, to Briana and Jenny, the emotional support and levity you have provided, especially in moments of despair, is incalculable; your friendship is one of the greatest gifts I have the honor of leaving UNC with.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1144081. The opportunities and resources afforded to me from this fellowship are immeasurable, as is my gratitude.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Advancements in many areas have contributed to the vast incorporation of precision medicine in current patient treatment plans. This includes, but is not limited to, genetically determining which biomarkers are associated with particular diseases (Jain 2002), increased efficiency of drug delivery systems (Allen and Cullis 2004), and more complex data collection procedures which provide richer data to make inference on (Cai et al. 2011). The goal of precision medicine is to incorporate a parsimonious amount personalized information to efficiently determine which treatments are best for which types of patients (Norvig et al. 2010, Hamburg and Collins 2010). This results in strategies that treat particular diseases by targeting a certain genetic marker, strategically controlling the amount and frequency of dosages and determining which patients should be receiving which treatments based on their demographic and clinical information.

A key component of precision medicine is creating mathematical estimators for clinical decision making. Statistical methodological research provides a objective way to predict which treatment, or which sequence of treatments, will lead to the best outcomes for each patient. While this type of work covers a wide variety of solutions to real world problems, there are always existing methods that can be improved upon or new areas to be explored. Many of the current optimal treatment estimation strategies are only designed with one outcome in mind. However, for a lot of patients, and a lot of therapeutic areas, it is more likely that many outcomes affect a patient's overall wellness. A patient may need to balance two failing organ systems, efficacy

and side effect burden, or even quality of life and cost. There are currently three primary approaches to estimation of optimal individualized decision rules for competing outcomes: (i) set-valued treatment regimes (Laber et al. 2014b, Lizotte and Laber 2016); (ii) inverse-preference elicitation (Lizotte et al. 2012a) ; and (iii) constrained estimation (Linn et al. 2016). Each of these either assumes a fixed composite outcome or does not address patient preferences directly through elicitation.

While the ideal situation is to directly elicit patient preferences, the type of elicitation where the patient chooses parameters to define a composite outcome is not feasible unless patients have undergone specialized training (Brennan 1998, Braziunas 2006, Lizotte et al. 2012a). Thus, the preferences must be indirectly estimated from information that can be directly obtained. A common approach for preference elicitation is to administer an itemized questionnaire that characterizes how the patient feels about each outcome in relation to the other. We assume the questionnaire is comprised of a series of binary responses to each question and use item response theory (Embretson and Reise 2013) to estimate the conditional distribution over these preferences. We use this conditional distribution to derive preference-sensitive optimal individualized treatment strategies for each patient.

This work proposes three methodological advancements that seek to solve the problems presented here. For these purposes, we only assume two competing outcomes. This research aims to develop rigorous, yet practical, ways to elicit and estimate a patient's latent preference information and incorporate it when making treatment decisions. This ensures that patients play a key role is decision making and offers a clear, logical way to link patient's preference and outcomes to create a well-defined utility function that serves as a composite outcome. The first method presented predicts single stage individualized treatment rules by estimating the latent preference information conditional on the itemized responses through the Rasch model (Rasch 1961;

1980). This preference information is standardized and incorporated into a linear utility function. The optimal treatment rule is that which optimizes the expected utility function for a given set of patient covariates. The second method extends this to the multiple stage scenario where the preference information is updated at each treatment stage based on how a patient's preferences have changed and their overall contentment with their health status. Using a reinforcement learning technique called Q-learning (Watkins and Dayan 1992), the optimal treatment rule is determined as that which provides the maximum expected utility given the patient's history, assuming that the best treatment is assigned at all subsequent stages. The final method provides a more flexible estimator by solving this problem nonparametrically. The preference information is conditioned on the itemized responses through constrained monotonic splines (Villalobos and Wahba 1987). To estimate the patient's ideal trade off between the two outcomes, we define a nonparametric utility function using monotonic splines once again to calibrate the patient's satisfaction with the outcomes after receiving the treatment. This more accurately captures the patient's feelings on how to weigh the two outcomes.

Obtaining data to evaluate this methodological research is difficult because the specific measures required are not collected during standard clinical trials. We are fortunate enough to obtain data from the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) schizophrenia trial (Stroup et al. 2003), which can be used in the single stage, linear model. Unfortunately, when this line of work was extended to two stages or incorporates a nonparametric estimator, the model assumes information that was not collected in this trial nor was found in any other trial of its kind. However, the set of up the CATIE trial can serve as a motivating example for the single stage, multistage or nonparametric paradigms. Schizophrenia is a chronic, severe and

debilitating mental disorder that may make patients lose touch with reality and can affects mental cognition, emotional stability and social behaviors (Andreasen and Flaum 1991). Since it cannot be cured at this time, treatments are designed solely to alleviate symptoms. One of the main treatments for schizophrenia is the administration of antipsychotic medications. Some of these antipsychotics elicit negative side effects in the patient, making it difficult for them to adhere to the treatment regimes, while others are less efficacious but also have less side effects (Lieberman et al. 2005). It is important to develop treatment plans that can find the appropriate balance between relieving schizophrenic symptoms and reducing the side effect burden. We can use this example to provide a realistic scenario when generating data for our simulation study. While this is one example of the application for this work, similar comparisons can be made with diseases such as rheumatoid arthritis and diabetes.

The remainder of this thesis is divided into the following chapters: chapter 2 contains a literature review that includes an introduction to item response theory and methods for single and multiple stage estimation; chapter 3 contains a method for incorporating patient preference when estimating a linear utility function in the single stage paradigm; chapter 4 extends this method to a multi-stage decision making process that incorporates evolving patient preferences and the patient's contentment with the results; chapter 5 provides a nonparametric approach to incorporating patient preferences for single stage estimation; and chapter 6 briefly summarizes the work presented and provides thoughtful areas of extension.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Personalized medicine is the practice of tailoring healthcare plans to patients as individualized as possible. In an ideal setting caregivers are able to account for patient history, genetic information, response to treatment, and other relevant factors when assigning each treatment throughout the entire course of the disease. These treatments would be assigned based on how each patient is progressing at each stage and which sequence of treatments is best earns the optimal expected long term response. These treatment plans have been coined dynamic treatment regimes (DTRs) and estimating DTRs is the main goal of personalized medicine. These treatment plans are unique in the sense that they are not predetermined at the initial doctor's visit as a blanket treatment for all patients. Instead, only the first treatment is decided at the initial visit (either a generic starting treatment or one determined by the patients/ baseline characteristics) and subsequent treatments are determined at each follow up period based on their progress. This makes the treatment regime dynamic through time. The evidence based treatment decisions are determined by the patients' response to the course of treatment and uses prognostic and treatment information to define their history (Collins et al. 2007). DTRs have four components: decision points (time points at which the decisions are made), tailoring variables (patients' prognostic information used to make treatment decisions), intervention components (the type or dose/intensity/duration of the treatment), and decision rules (a function that links the

tailoring variables to the intervention options at decision points) (Collins et al. 2014).

In the statistical framework, DTRs are developed to create mathematically dependent treatment plans that mimic how clinicians treat patients in practical settings. This quantitative way of estimating treatment plans is a new and novel way to think of treating patients. Preliminary interest in this way of thinking began in the mid-1980s when researchers were evaluating two stage dynamic treatment strategies for cancer patients. In the 1990s the idea of a two stage customized treatment plan expanded when psychiatrists were interested in expanding to k-stage treatment plans. By the early 2000s, clinical trials and treatment strategies were implemented by doctors studying DTRs in substance abuse and mental health research (Lavori and Dawson 2014). Once the physician or researcher has conjured a list of potential DTRs, they need to be evaluated in a special clinical trial called a Sequential Multiple Assignment Randomization Trials (SMART) (Murphy 2005a). The development of SMARTs began when traditional clinical trial designers were looking to identify an intermediate outcome between randomization and the primary outcome and make an appropriate reaction to this outcome. The goal was to be able to assess the status of the patient throughout the trial and adjust the treatment as necessary.

It is important to highlight the distinction between DTRs and SMARTs. A DTR is a treatment strategy that tells the caregiver which (sequential) treatment plan will lead to the highest probability of success and adapts to the patient's health status throughout the course of treatment. A SMART design is an experimental trial with the purpose of determining which DTRs are optimal for which patients. A SMART is able to evaluate the various DTRs because it compares between them by randomizing patients to preselected treatments. A SMART can have two goals: compare a small number of prespecified DTRs embedded in the SMART design or construct new DTRs which may not be naturally embedded in the design.

The material presented in this literature review is meant to be a survey of current statistical methods used to estimate dynamic treatment regimes in the context of using data collected in a SMART trial. Chakraborty and Moodie (2013) have published a book titled "Statistical Methods for Dynamic Treatment Regimes" which covers this and similar topics in some depth. In this book, the authors discuss select topics such as SMART designs, observational studies, reinforcement learning and various methods of estimating DTRs. Please refer to this book for an indepth description of a number of useful methods for DTRs.

This review covers a brief introduction to item response theory, how to design a SMART and what statistical methodologies exist to make inference from this data, with particular emphasis on recent developments in the area. Section 2.2 introduces item response theory and the Rasch model. Section 2.3 will highlight how a SMART design fits into the overall development of a treatment strategy, the importance of pilot studies, practical considerations such as sample size, calculations, and handling missing data, as well as the future of data collection. Section 2.4 describes select methods for choosing tailoring variables. Section 2.5 introduces methodological results for the single stage paradigm and section 2.6 provides methodological results for the multiple stage paradigm. Section 2.7 describes methods designed to develop DTRs when the only data available is observational because a randomized design is not practical, is impossible, or is unavailable. Section 2.8 describes the work done developing DTRs for competing outcomes.

2.2 Item Response Theory

An important piece to this research is incorporating patient's preferences into their treatment decisions. To do this, a latent trait model from item response theory

will be employed to estimate a subject's implicit utility function regarding the two competing outcomes so that it can be included in estimation. This is done by estimating parameters from the Rasch model that is conditional on a unidimenional latent trait.

Item response theory (IRT) concerns the accuracy and development of test scoring when these tests or questionnaires are designed to measure abilities, traits, or behaviors (An and Yung 2014). The test or questionnaire consists of a set of items (questions) with binary or ordinal responses. These models not only provide accuracy of test scoring, but can also improve efficiency of data collection by only including the significant items. Historically, IRT has been primarily used for psychological assessments and educational testing, but these models have recently been extended to health research. This is partially because of the aforementioned ability to tease out only the discriminative items for inclusion. Outside of the education and psychological areas, IRT is referred to as Latent Trait Modeling (LTM).

Formally, LTM is latent structure analysis of categorical data and can be thought of as factor analysis for dichotomous or ordinal data (Uebersax 2000). To first understand this idea, it is easiest to consider the model in terms of its latent structure. This modeling technique reduces the set of dichotomous variables to a set of small factors called latent traits. Their desirability stems from the fact they are formalized probability models which relates the unobserved (latent) variables to the observed variables that are measurable in real life. These latent trait models allow precise measurement of the difficulty of items, determines the association of each item with a person's proficiency, determines which items are biased for different subpopulations, creates questionnaires with the minimum number of items, measures accuracy at different levels of proficiency, and allows for the ability to create adaptive tests where preceding items determine the subsequent ones. From here on out, although interchangeable we will refer to both LTM and IRT as just IRT. Within each line of thinking there are minor differences,

but computationally they are essentially the same.

IRT is advantageous over classical test theory (CTT), one of the original methods of test response analysis, because IRT takes both information from the person and the items into account (Yu 2013). CTT cannot differentiate between the subjects proficiency and the difficulty of the item; meaning for populations of different proficiencies, the questionnaires could seem easy or difficult (An and Yung 2014). On the other hand, IRT assesses both the difficulty of the question and the subject's proficiency throughout the entire questionnaire. Once the items are calibrated for the population, scores for subjects can be compared even if they did not answer the same set of questions. This calibration requires an iterative process because the proficiency and difficulty determined by the data are used to fit the model which in turn predicts the data. The ability to compare subjects who respond to different sets of items also reduces the size of the questionnaire. This increases efficiency and in turn reliability because the precision differs based on latent structure and can be generalized to the entire population. IRT modeling is considered a superior analysis method and is employed in many standardized exams, such as the Scholastic Assessment Test (SAT) and the Graduate Record Examination (GRE) (An and Yung 2014).

One of the most well known models used in IRT is the Rasch Model (RM) (Rasch 1980). Before continuing, it is important to point out the philosophical differences between IRT modeling and Rasch modeling even though computationally they are similar. Item response theory purists build models with the intention of creating a model that fits reality, which contains a lot of parameters and becomes very complicated. Here, the goal is to fit the model to the data. Rasch modelers seek to build elegant, simple models with more practical implications and fit the data to the model (Yu 2013). The Rasch model is a model employed by item response theorists to analyze itemized response data.

The one parameter Rasch model is designed for dichotomized response data, i.e $\{1, 0\}$ where 1 corresponds to “yes" and 0 corresponds to “no". The probability of responding “yes" is:

$$P(x_{ij} = 1|\theta_i, \beta_j) = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}}$$

where $P(x_ij = 1|\theta_i, \beta_j)$ is the probability of person i responding “yes" (1) to question j (Li and Baron 2012). The latent trait, θ_i, is thought of as person $i's$ proficiency and β_j represents the difficulty of item j. The probability of answering yes to each item is dependent on the person's proficiency and item's difficulty such that if the person's proficiency matches the item's difficulty, that person has a 50% chance of answering yes.

The Rasch model can be extended to a two parameter model by introducing a discrimination parameter, which measures the slope, to allow more flexibility in modeling (An and Yung 2014). In this model, the probability of responding “yes" is:

$$P(x_{ij} = 1|\theta_i, \beta_j) = \frac{e^{\alpha_j\theta_i - \beta_j}}{1 + e^{\alpha_j\theta_i - \beta_j}}$$

where the discrimination parameter, α_j, measures if the item has the ability to differentiate subjects. A high discrimination parameter (α_j) implies that the probability of answering "yes" increases more rapidly as the proficiency parameter (β_j) increases. This parameter tells how effectively the item discriminates between highly and lowly proficient students (Yu 2013)

These two models are utilized for dichotomous responses, but multiple extensions of these models have been developed for increased flexibility. These include, but are

not limited to, three and four parameter models, Bayesian IRT models, models used to analyze ordinal responses (graded response models), and models used to analyze items that can be explained by more than one latent trait (multidimensional models).

Letting $\pi_{ij} = P(x_{ij} = 1|\theta_i, \beta_j)$, the Rasch model can easily be transformed into a logistic regression model (or a probit model) such that:

$$logit(\pi_{ij}) = \alpha_j\theta_i - \beta_j$$

where θ_i is the latent variable and α_j, β_j are parameters that can be estimated from the data.

2.3 Study Design

A general framework used to create a clinical treatment strategy is the Multiple Optimization Strategy (MOST) (Collins et al. 2005). The MOST is an engineering inspired framework which consists of 3 phases. First is the screening or preparation phase where the intervention components are identified for inclusion or rejection. Treatment or delivery methods are also chosen at this time based on theoretical assumptions determined by the physician. Previously garnered information is used to create a guide for the selection of intervention components, which questions need to be answered and what outcome is being optimized. Second is the refining or optimization phase, where the selection components are tuned and optimized with the goal of arriving at a final proposal of the treatment protocol. The optimization phase is information gathering and decides which treatment combinations achieve the optimization criterion via a randomized experimental design that proposes a sequential treatment intervention. Last is a confirmation or evaluation phase, where the optimized sequential intervention is

evaluated in a traditional randomized controlled trial (RCT) where they investigate efficiency and practicality. At times the first two phases need to be repeated before the final stage can be implemented. The MOST assumes that experimentation makes the most efficient use of available resources with the goal of producing the largest improvement and treats optimization as a process rather than an endpoint.

Traditionally, the screening/preparation and refining/optimization phases used a factorial or fully crossed analysis of variance (ANOVA) design. Because SMART designs are considered a special case of the factorial design where not all factors need to be crossed, they have been found to be advantageous to replace the ANOVA model in the second phase. A SMART creates the adaptive intervention by allowing for dynamic treatments (as opposed to fixed ones) and provides a basis to identify the best tailoring variables and decision rules. Recall that tailoring variables are essentially patients' prognostic information used to make treatment decisions and the decision rule is a way of choosing which treatment to assign. They investigate the best sequencing of intervention components, what tailoring variables should be used, when and how frequently should these tailoring variables be assessed, and should one treatment be assigned or should patients have the ability to choose from a list of options. The treatments are assessed as an entire treatment sequence and not isolated by phase of treatment. They involve multiple randomizations over time, where each randomization point corresponds to a decision point and questions are investigated regarding two or more treatment options at each decision point. Additionally, responders and non-responders are controlled for by design because different treatment decisions need to be considered for each (Collins et al. 2007; 2014).

Consider a generic example for illustration. A 2 stage SMART enrolls 200 participants with equal randomization such that it assigns 100 patients to treatment A

and 100 patients to treatment B at stage 1. At the end of stage 1 the patients' response to the treatment is assessed using the response variable or tailoring variables, and the patients are classified as either responsive or nonresponsive. What defines responsive or nonresponsive should be determined a priori. Each of the patients is then re-randomized according to their response classification. For instance, patients who responded to treatment A are assigned to stay on treatment A at stage 2, while the patients who did not respond to treatment A are randomized to treatment C or D. A similar situation could arise for treatment B where responders stay on treatment B and nonresponders are randomized to treatments E or F. In this scenario, there are 6 DTRs: A-A A-C A-D B,B B,E B,F. Alternatively, responsive patients could also be re-randomized. There could be more than 2 randomization options at each stage, or the nonresponsive patients could switch from A to B or vice versa. The randomization also does not need to be balanced, and the stages could be extended beyond only 2. In this example data would be collected at 3 time points: baseline, time 1 (after the first stage of the study but before the patient is re-randomized to the second treatment), and time 2 (after the second stage or at the end of the study). Generally speaking, randomization does not need to depend on responder status, although this is the case in this example.

There are numerous reasons it is advantageous to use a SMART. For example, it has the desirable quality of making better use of the pre-determined sample size and can answer more clinical questions than a RCT. In a two phase trial, the first phase of the trial can be assessed by comparing the mean outcomes between the first two lines of treatment. The second phase can compare the effect of the treatment options for responders and nonresponders, regardless of their first line of treatment. This increases the power of the hypothesis tests, since they recycle the patients in the second phase of the study. Most importantly, the design then has the ability

to compare the embedded treatment regimes that pool information across multiple experimental conditions by, again, recycling patients through the trial (Collins et al. 2014). While the main goal of a SMART is to mine data used to develop DTRs, they have other uses, such as discovering which treatments work best sequentially to obtain an improved outcome, investigating the interplay between trajectories of the patients disease progression and treatment sequences, comparing different treatment sequences, and investigating the benefit of both prognostic information and observable data in determining individualized treatments (Almirall et al. 2012).

Now that it is clear what a SMART is, why it is beneficial, and what research would benefit from it, what does one do next? How does one begin to design a SMART? The answer, as in most clinical trials, lies with running a pilot study. When researching any new treatment or treatment strategy, a pilot study is always advised, if not required, and this is no different. A good pilot study is great practice for implementing a larger design, gives critical value estimates, is important for sample size estimation, and provides a preliminary look at the utility of the proposed treatments. The novelty of SMARTs raises feasibility concerns which makes a pilot study even more crucial for designing an effective and efficient SMART. Note that a SMART pilot is usually performed before the first phase of the MOST because it provides information required for justification of a theoretical model. Almirall et al. (2012) highlights important topics that researchers must be aware of when designing a SMART and how solutions can be elicited through a pilot study: (1) considering the choice of a primary tailoring variable. This will be used to determine the set of randomized treatment options such as assessing response/non-response, when this decision should be made, what criterion is used to make the decisions, frequency of assessment, how sensitive this measure is, justification of its use, and feasible application in clinical treatment; (2) deciding which

additional potential tailoring variables should be collected, such as either baseline patient characteristics or time varying measures. These are determined as important in predicting outcomes to later stage treatments; (3) deciding how to control for missing tailoring variables. This should be guided by how it would be handled in clinical practice; (4) deciding between up-front randomization (randomization at the beginning of the trial) or real-time randomization (randomized sequentially at each decision point which allows for clinical information to be used in randomization); (5) highlighting the difference between research assessments for data analysis to develop adaptive treatment strategies and assessments of the adaptive treatment strategies used to inform the sequential treatment assessment; (6) identifying concerns clinicians have regarding sequences of treatments offered and assessment of what determines response versus nonresponse; (7) assessing for patient acceptability; (8) testing the language of consent forms; (9) illuminating unanticipated tailoring variables that would be useful in the subsequent SMART.

The pilot study provides crucial information to increase the probability of success of a SMART. However, even with a pilot study, often times there is still not enough information collected to prevent unanticipated hiccups. As previously stated, one goal of a SMART is to identify embedded DTRs not embedded in the original design. A great example of this is the analysis of SMART data collected from a study for the treatment of advanced prostate cancer. Wang et al. (2012) created a new method to compare dynamic treatment regimes and along the way, changed the definition of the DTRs after the trial ended. This analysis is different than previous analyses using the same data because they changed the definition of viable DTRs based on what had been predetermined as a 'missing observation'. The protocol required that patients were re-randomized to a new treatment and if they did not complete it, they were classified as missing. However, the treatment plan determined by the protocol was not

feasible for patients with toxicity or disease progression. They altered one of the DTRs to include those patients who had to leave the study because of toxicity where the second treatment was the recovery treatment they were given after leaving the trial. The viable DTRs were now defined by efficacy, toxicity and disease progression. The authors also redefined this endpoint because it needed to quantify the health experience of the patient over a pre-specified fixed period, not just their final tumor size or toxicity.

By now it should be clear that designing, implementing and analyzing SMARTs is a relatively new area, which means there is still a lot of work to be done. While there is a lot of methodological work needed or expanded upon (see subsequent sections), there are still a lot of gaps in knowledge for designing these trials. One clear gap in the literature is work on a universal (or adaptable) sample size formula. There are sample size calculation publications for SMART designs, but there is no universally accepted calculation for general use.

Fortunately, there has been some progress made in sample size formulas for SMART design, such as the development of upper bound sample size estimates for censored data. Unfortunately, this sample size formula does not have great generalizability properties because the upper bounds are based on a Kaplan-Meier estimate and the log-rank statistic. Li and Murphy (2011) developed a sample size formula for a two stage randomized trial (with the goal of developing DTRs) for failure time outcomes. The difficulty in such a calculation stems from the variances of the common test statistics. These test statistics depend on the joint distribution of the time, early response determination, and the primary failure time, which are likely to be dependent. The sample size is derived using upper bounds on the variances in place of the usual variances and, hence, the resulting formula only requires the same assumptions of a traditional single stage randomized clinical trial. Using the upper bounds of the variances, the proposed samples size formulas for the Kaplan-Meier estimate (n_K) and

the log rank statistic (n_L) are

$$n_K \leq \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 \sigma_B^2}{\left\{\bar{F}_1(\tau) - \bar{F}_2(\tau)\right\}^2}$$

and

$$n_L \leq \left(\frac{1}{pq} + \frac{1}{(1-p)q}\right) \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\xi^2 \int_0^t \bar{F}_c(t) \mathrm{d}F_1(t)}$$

respectively, where Z is the z-score for a standard normal distribution, α is the type I error, β is the type II error, $\bar{F}$ is the survival function, τ is the time at the end of the study, $p = p(A_1 = 1)$ and $q = p(A_2 = 1|R = 1)$ are the randomization probabilities, R is the indicator of randomization, ξ is the log hazard ratio. Note that

$$\sigma_B^2 = \frac{\bar{F}_1^2(\tau)}{pq} \int_0^\tau \frac{\mathrm{d}\Lambda_1(t)}{\bar{F}_1(t)\bar{F}_c(t)} + \frac{\bar{F}_2^2(\tau)}{(1-p)q} \int_0^\tau \frac{\mathrm{d}\Lambda_2(t)}{\bar{F}_2(t)\bar{F}_c(t)}$$

where Λ is the cumulative hazard function. This sample size calculation has been proven to provide the desired power if the hazards of the alternative are proportional. This sample size calculation is most notable because the nature of chronic diseases (the focus of SMARTs and DTRs) allows the outcome, or surrogate outcome, to be thought of in terms of failure, even if death is not the primary endpoint. This means that this sample size formula will be applicable in many settings, especially until further research is done.

Another clear gap in the literature is lack of progress made in developing strategies that prevent, and methodology that controls, for missing data. The construct of a SMART requires numerous randomizations and multiple treatment prescriptions which

presents unique challenges when analyzing data in the presence of missing data. Imputation strategies for handling missing data collected from SMARTs is an understudied area at this time. Shortreed et al. (2014) presented the following five missing data issues: (1) transition between treatment stages does not always occur at pre-specified times but instead can be determined by a patient outcome; (2) some outcome variables are irregularly spaced while some variables are collected at regularly scheduled study visits; (3) observing some variables is dependent on a patient's history, which results in structural missingness for the data-dependent portion of the collected information. (4) individuals are simply lost to follow up leaving the treatment stage; (5) some individuals are lost to follow up entering the treatment stage. Their proposed solution is a flexible imputation strategy to facilitate valid inference using data from SMARTs which is a time ordered, nested, conditional imputation strategy, which exploits the nearly monotone pattern of missing data found in this type of longitudinal study. It ensures that a complete multivariate prediction distribution exists while obtaining desirable traits for inference across longitudinal outcomes. Assuming missingness at random, this method works best when the data is imputed with a pseudo-Gibbs sampler, which applies repeated iterations through the model. Multiple imputation is one of many strategies used when working with missing data, and the type of strategy often depends on the structure of the data and the nature of the missingness. This method was not compared to other imputation strategies such as inverse probability weighting or likelihood methods, and there is no contingency plan when the missingness is not monotone. It is clear the presented work is exciting and promising progress, but more headway is still needed.

While momentum in some necessary elements of SMART study design is stalled, other areas are moving full speed ahead. An exciting area of expansion is data collection and treatment allocation using mobile technology. The clear interest is the ability to

increase access to fast accurate care because mobile technologies include but are not limited to cell phones, sensors and monitors. The goal is development of evidence based Just in Time Adaptive Interventions (JITAIs) that collects real time data from patients and uses that data to inform the real time delivery of intervention options, such as treatment, dose and timing of care (Nahum-Shani 2013). For example, trying to intervene in heavy drinking and smoking, a mobile phone would be administered and participants would be prompted 3 times a day to assess their smoking urge, affect, and drinking behaviors. Urge management interventions would be delivered only if the individual reports the urge to smoke at a specific time. Anytime during the day the user can text either lapse or crave, and a series of encouraging text messages will be sent back to their cell phone. Another example is managing eating disorders. When treating college women with eating disorders, the subject would be provided with a cell phone which receives 5 prompts regarding mood, eating behaviors, exposure, etc. When she reports what is considered a negative mood she is recommended to use one of the treatments provided via a CD. In all three instances, the interventions are adapted and delivered through a mobile medium such that patient information can be obtained at any time and responses can be administered at any time. The variety of potential interventions includes reach out interventions, behavioral strategies, cognitive strategies, and goal setting. Tailoring variables can be collected actively (self-reported via prompting or user initiated) or passively (activity level, location, social media activities, number of ignored recommended interventions, etc.). The decision points vary depending on the goal of the treatment plan and can include a random prompt, user requested help, or indication of specific experiences. The decision rules can be deterministic (if the patient reports more than X then give them this, otherwise give them that) or stochastic (determining the probability of an intervention). The corresponding thresholds can be determined and optimized using reinforcement learning

(which will be covered in later sections). Even though this concept is in the early stages of development, there are obvious feasibility issues for this kind of treatment implementation, such as cost and monitoring adherence. Innovative methods like this have the ability to change the way patients are treated and can serve as a guide for future treatment and collection methods.

A SMART is a novel approach to efficiently collect data which accurately estimates DTRs for individualized patients. The basic design structure has been created, trials are ongoing in the clinical setting, and new advances are being developed day by day, but more work needs to be done. The flexibility of the design makes developing broad techniques difficult, but the need and the talent is there to continue to make advancements in what has become an extremely timely, interesting and applicable area of study.

2.4 Variable Selection

Variable selection is an important component of estimating optimal DTRs because tailoring variables are used to adapt the treatment plan to the individual. The goal is to avoid a priori hand picking tailoring variables, but instead use the data to select a subset of the tailoring variables that estimates a decision rule as close to the optimal rule estimated when using all variables. Including all possible variables as tailoring variables is inefficient and will often lead to over fitting. Once the tailoring variables are selected, they can be used when optimizing DTRs. A brief overview of recent and relevant variable selection techniques is included in this section.

Biernot and Moodie (2010) discuss two computer science techniques that can be used for variable selection: the S-score criterion and the use of reducts. The S-score of a variable shows the expected increase in response that is observed by choosing

the treatment based on the value of that variable. It combines the interaction of the covariate with the treatment and the proportion of the population exhibiting variability in that covariate. Higher values indicate stronger relationships between the variable and the treatment, and shows that a large proportion of patients would experience change in the optimal action if the variable was taken into consideration. This scoring is used to rank potential variables but each variable is evaluated separately meaning correlation between variables is not taken into consideration. The S-score could also be used sequentiality such that the variable with the highest score is first selected, then the variable with the second highest score given the first variable is selected and so on.

The reducts approach was developed from rough set theory in computer science. The positive region is a set of all observations that can be uniquely classified into one equivalence class based on the non-decision variables. The reduct is the minimal set of tailoring variables that classifies individuals into unique decision equivalence classes as well as the complete set of variables does. Reducts help eliminate redundant variables while preserving information regarding the similarity of individuals in the sample. In the scenario with multiple reducts, one can select the variables most frequently seen in the reducts, or can select amongst reducts by choosing the set of covariates with the highest S-score. This last hybrid method is believed to combine the strengths of these two methods. Unfortunately, it is important to note that reducts are not appropriate for continuous outcomes.

Another way to approach variable selection is to simultaneously estimate optimal treatment regimes and identify significant variables. This is done with a penalized regression model that finds which variables interact with the treatment using a new loss based framework. Lu et al. (2013) introduces a method which does not require estimating the baseline mean function for the outcome of interest and is easily adaptable to shrinkage methods for variable selection based on their loss structure making it

quickly implementable with current software. The authors suggest the loss function

$$L_{n,\phi}(\beta,\gamma) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \phi(X_i;\gamma) - \beta^T\tilde{X}_i\{A_i - \alpha(X_i)\}\right]^2$$

where n is the number of observations, Y_i is the i^{th} patient's outcome, X_i is the i^{th} patient's prognostic variables, $\tilde{X} = (1, X^T)^T$, $A_i \in \{-1, 1\}$ represents the dichotomous treatment choice, $\alpha(x)$ denotes the propensity score, and ϕ is an arbitrary function with a constant model for $\phi : \phi(x;\gamma) = \gamma$ and a linear model for $\phi : \phi(x;\gamma) = \gamma^T\tilde{x}$. This characterization of the loss function increases simplicity in adopting shrinkage penalties for variable selection. Employing the adaptive lasso penalty (or, alternatively, the SCAD or minimax concavity penalty) the solution is the β which satisfies

$$\min_{\beta} L_{n,\phi}(\beta,\tilde{\gamma}) + \lambda_n\sum_{j=1}^{p+1} w_j|\beta_j|$$

where λ_n is a tuning parameter and w_j are the weights such that $w_j^{-1} = |\tilde{\beta}_j|$ is used. Aside from estimating the optimal DTR, these β values are used to determine which variables are important in selecting the optimal DTR such that the important variables are those with nonzero coefficients.

Variable selection is an important part of estimating optimal DTRs because a parsimonious selection of tailoring variables will make the estimation faster and more reliable. Three methods have been presented here for these purposes, but more methodology has been published. It is imperative to use a selection technique that is relevant for the data set and can be effectively integrated into the analysis plan.

2.5 Analysis Techniques: Single Stage

The list of methodology presented here and in the subsequent section is neither complete nor representative of all available options, but a simple summary of recent or relatively recent methods employed along a broad range. The purpose is to introduce popular techniques, highlight advancements, and display a plethora of methodological options applicable in multiple areas of interest.

Important notation must be introduced so that an individualized treatment rule (ITR) for the single stage paradigm can be properly defined. An ITR differs from a DTR in that it is the personalized rule for a single treatment setting while a DTR is the sequence of decision rules for a multiple treatment setting. Assuming the data is collected from a single stage two arm trial, the treatments will be annotated as $A \in \{-1, 1\}$. These are independent of the patient prognostic variables denoted as $X = (X_1, \ldots, X_p)^T$, where X is a p-dimensional matrix. In the single stage paradigm the observed clinical outcome, Y, can be considered a reward function where larger values are desired. The ITR is a map from the prognostic variable space, X, to the treatment space, A, and the optimal ITR is the A which maximizes the expected reward. The distribution of (X, A, Y) is denoted by P with the respective expectation denoted as E. The distribution of (X, A, Y) given the ITR, D (i.e. that $A = D(X)$), is denoted as P^D and the corresponding expectation as E^D. The expected reward under D is

$$V(D) = E^D(Y) = E\left[\frac{I\{A = D(X)\}}{A\pi + \frac{1-A}{2}}Y\right]$$

where $\pi = P(A = 1)$. This $V(D)$ is referred to as the value function for a given D. The optimal ITR, denoted D^*, is estimated as:

$$D^* \in argmax_D V(D) = argmax_D E\left[\frac{I\{A = D(X)\}}{A\pi + \frac{1-A}{2}} Y\right]$$

and is considered the D which maximizes the value function $V(D)$. The optimal treatment regime is defined as the one that maximizes the average expected outcome (Zhao et al. 2012).

One way to estimate ITRs is to restructure the estimation procedure into a classification problem where the optimal classifier corresponds to the optimal treatment decision. The optimal classifier can be found by estimating the Bayes classifier, which is the one that minimizes the expected weighted misclassification error. This framework allows for estimation of mean outcomes under existing methods such as regression estimation, inverse probability weighted estimation (IPWE) or augmented inverse probability weighted estimation (AIPWE) (Zhang et al. 2012a). The class of treatment decisions is data driven because it is chosen by minimizing the L1 the expected weighted misclassification error and does not need to be prespecified.

Define the contrast function as

$$C(X) = \mu(1, X) - \mu(-1, X)$$

which can be thought of as the mean difference between treatment options for a given set of prognostic variables. The optimal ITR estimation problem can be transformed into a weighted classification problem such that the optimal treatment rule D^* is found by

$$D^* = argmax_D E\left[D(X)C(X)\right] = argmax_D E\left(|C(X)|\left[I\left\{C(X) > 0\right\} - D(X)\right]^2\right)$$

This means that the optimal treatment rule, D^*, is found to be the one that maximizes $E\left(|C(X)|\left[I\left\{C(X) > 0\right\} - D(X)\right]^2\right)$, which is a weighted classification problem. Each subject belongs to two classes such that class $Z = 1$ contains those subjects who would benefit more from treatment $A = 1$ as opposed to treatment $A = -1$, e.g. $\mu(1, X) > \mu(-1, X)$, and $Z = 0$ the opposite. Each observation is also given a weight, $W = |C(X)|$, which is the loss that would incur from misclassification. Hence, the optimal ITR is simply the expected weighted misclassification error under the classification rule $D(X)$. Within this classification construct, the problem then decomposes into two critical steps. First, one must construct a suitable estimator of the contrast function, using regression, and then invert this to find the estimated optimal treatment rules with an interpretable form using classification methods. This can be extended to the multiple stage scenario as well. This classification prospective falls under the machine learning umbrella. Machine learning, most specifically reinforcement learning, has recently been implemented since it sidesteps the problem of completely modeling the underlying generation model as is necessary in some estimation techniques. Reinforcement learning is a dynamic programming system that decides which actions need to be taken to optimize a given reward.

Qian and Murphy (2011) propose a modification of this which first estimates the conditional mean response using l_1 penalized least squares (l_1-PLS) with a rich linear model and then uses that to derive the estimated treatment rule. If the conditional mean is modelled correctly, this method consistently estimates the optimal treatment rule. The finite sample upper bounds of the difference between the mean response

from the optimal treatment rule and the mean response from the estimated treatment rule holds even if the linear model for the conditional mean response is incorrect. If the part of the conditional mean model involving the treatment effect is correct then the upper bounds imply that the estimated treatment rule is consistent. These upper bounds can also inform how to choose the tuning parameters involved in the l_1-penalty to create the best rate of convergence. To obtain the ITR the estimated prediction error is minimized then the conditional mean model is maximized over the treatment A. To control for overfitting, l_1 penalized least squares is implemented since the l_1 penalty innately does variable selection. The resulting treatment rules are cheaper to implement and easier to interpret.

The forgoing methods are considered indirect methods of estimation. Indirect estimation refers to techniques that first estimate a quantity reflecting the conditional distribution of the outcome, such as conditional mean, and then uses the resulting model to deduce the optimal ITR (Laber et al. 2014c). Indirect estimation can be desirable because the initial estimation regarding the outcome can be built using traditional statistical modelling techniques. Unfortunately, optimal ITR estimation requires that the conditional outcome be modelled correctly. Indirect methods of estimation often and easily experience model misspecification because of the difficulty of modelling high dimensional, time dependent factors. In high dimensional situations the two-step procedure of estimation and maximization equations can be poor fits. In contrast, direct methods of estimation are solutions to the problems proposed by these other techniques which directly achieves this maximization without requiring the initial estimation step be done with indirect approaches. This direct class of methods immediately estimates the value function for all prespecified treatment rules and then obtains the optimal treatment rule by maximizing the estimator. Direct estimation methods tend to produce treatment regime estimates that are more precise than indirect methods in the

single stage setting due to the associated reduced bias (Zhao et al. 2012).

Zhang et al. (2012b) approaches estimating dynamic treatment rules by assuming a posited regression model. This defines the class of treatment rules while recognizing that it is possible for the model to be misspecified. The optimal treatment regime is estimated by directly maximizing the estimator for the overall population mean outcome under all possible specified treatment plans using a suitable inverse probability weighted estimator. When using observational data, this estimator has the ability to control for possible confounders by estimating propensity scores and exploiting the predicted outcome, which ensures precision of the estimate. Let D^* be the optimal treatment decision, which is the one that corresponds to the largest value of $E[Y^*(D)]$, where

$$Y^*(D) = Y^*(1)D(X) + Y^*(-1)\{1 - D(X)\}$$

is the potential outcome. The potential outcome is the outcome that would be observed if a randomly chosen patient were to receive treatment regime D. Consider treatment rules of the form $D_\eta(X) = D(X, \eta)$ in the class of all possible treatment rules which is indexed by η and will contain D^* if $\mu(A, X; \beta)$ the posited regression model is correctly specified. Therefore, estimating $\eta^* = argmax_\eta E[Y^*(D_\eta)]$ and defining $D^*_\eta = D(X, \eta^*)$ will provide an estimator for D^*. To estimate $E[Y^*(D_\eta)]$, an IPWE or a doubly robust AIPWE can be employed. This estimator is directly maximized in η to obtain an η^* and hence $\hat{D}^*_\eta(X) = D(X, \hat{\eta}^*)$. This can easily be extended to the multiple decision situation by estimating $Q(\eta) = E[Y^*(D_\eta)]$ as a function of η.

Direct methods of estimation can also be restructured into a classification problem which can utilize computer science techniques. This looks at the data by comparing the

difference between subjects with observed high and low rewards so that the determination of the actual treatment decisions is associated with the actual treatments received for the different groups. This method is referred to as outcome weighted learning (OWL or O-learning). Developed by Zhao et al. (2012), O-learning is a nonparametric approach which directly optimizes the value function $V(D)$ where each subject is weighted proportional to their clinical outcome divided by the propensity score, which is the probability of receiving the assigned treatment given the covariates. In the case of a clinical trial, the propensity simplifies to the constant probably of receiving the assigned treatment. Finding the D^* that maximizes $V(D) = E\left[\frac{I\{A=D(X)\}}{A\pi+\frac{1-A}{2}}Y\right]$ is equivalent to finding the D^* that minimizes $\bar{V}(D) = E\left[\frac{I\{A\neq D(X)\}}{A\pi+\frac{1-A}{2}}Y\right]$ which sets the stage to view this as a weighted classification error. Minimizing the previous expected value can be approximated by minimizing

$$n^{-1}\sum_{i=1}^{n}\frac{Y_i}{A_i\pi+\frac{1-A_i}{2}}I\left[A_i \neq sign\left\{f(X_i)\right\}\right]$$

to find the optimal f^* and then setting

$$D^*(x) = sign\left\{f^*(x)\right\}$$

since $D^*(x)$ can always be represented as $sign\left\{f^*(x)\right\}$. This implies the goal is to find a decision rule which chooses treatments based on their specific prognostic variables. On average, patients with large rewards will be recommended the same treatment that they actually received while patients with small rewards will receive the opposite. This is considered 0-1 loss in the machine learning scenario and is difficult to minimize due to non-convexity and discontinuity. This problem is alleviated by transforming the problem, using a surrogate for the 0-1 loss, so that the goal becomes minimizing

$$n^{-1}\sum_{i=1}^{n}\frac{Y_i}{A_i\pi+\frac{1-A_i}{2}}\left\{1-A_i f(x)\right\}^{+}+\lambda\|f\|^2$$

where $x^{+} = max(x, 0)$, and $\|f\|$ is the norm of f. Therefore, this problem is now a weighted classification problem that can be solved using support vector machine methods.

O-learning has many applications but there are also many ways to extend this line of thinking. Chen et al. (2016) presented a one stage clinical trial design for penalized dose finding using a robust analysis method based on the O-learning framework. The method converts the individualized dose selection problem into a penalized weighted regression with truncated l_1 loss. The dose level is assumed to be found on a continuum and a non-trivial extension of O-learning for binary treatments is proposed. The dose finding problem becomes a weighted regression with random outcome where the individual responses are the weights. In the linear case, this framework has the goal of minimizing the loss plus penalty of the form

$$\min_{f}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{R_i l_\phi\left\{A_i-f(X_i)\right\}}{2\phi_n p(A_i|X_i)}+\lambda_n\left\|f\right\|^2\right\}$$

where $\phi = \phi_n$ is non-random parameter in real space, λ_n controls the severity of the penalty on f, $l_\phi\left\{A_i - f(X_i)\right\} = \min\left(\frac{|A-f(X)|}{\phi}, 1\right)$, $R^*(a)$ is the potential outcome and $R = \int I(A = a)R^*(a)p(a|x)\mathrm{d}a$. The complexity of $f(x)$ is penalized to prevent overfitting. This function is nonconvex and hence difficult to optimize, so an adaptive difference convex (DC) algorithm is implemented (Tao and An 1997). Considering a linear loss function, the objective function is

$$S = \frac{\lambda_n}{2}\|w\|_2^2 + \frac{1}{\phi_n}\sum_{i=1}^{n} R_i \min\left(\frac{|A_i = D(X_i)|}{\phi_n}, 1\right)$$

where λ_n is now the tuning parameter. This algorithm minimizes the sequence of convex sub-problems with the intent of solving the original non-convex minimization problem. Therefore, the convex sub-problem becomes a weighted penalized median regression problem. Ultimately, the algorithm concludes when $\|w^{t+1} - w^t\|$ is smaller than some prespecified constant, where $w = \sum_{i \epsilon T}(\alpha_i - \bar{\alpha}_i)x_i$. Expanding to the nonlinear framework, the decision function then becomes a function of w and some unknown transformation on X. A Gaussian kernel is used to construct a dual problem for nonlinear learning that is solved using quadratic programming. To practically implement this procedure, a nonconvex loss function and a DC algorithm for optimization is employed.

Another common and extremely relevant application of O-learning is estimating ITRs for censored data. Realistically many chronic diseases measure short term success of a treatment as a failure or success. It is desirable to develop methods of estimating treatment regimes that are applicable to survival analysis because it has clear relevance to personalized medicine. When considering censored data, notation is slightly altered. The value function is redefined as

$$V(D) = E^D(T) = E\big(T|X, A = D(X)\big) = E\left[\frac{I\{A = D(X)\}}{A\pi + \frac{1-A}{2}}T\right]$$

where $T = min(\tau, \tilde{T})$ where $\tilde{T}$ is the survival time and τ is the end of the study (Zhao et al. 2015). Even though the outcome is redefined, the optimal treatment rule is still the treatment rule which maximizes the value function. The goal is to estimate D^* using censored data following the OWL framework. There are two approaches to estimation. First, one can maximize the estimator of the average survival time. To account

for right censoring, the estimated mean survival time is reassigned as the weighted misclassification rate. These weights are comprised of both the observed outcome and the inverse probability of censored weights. To offset bias from a misspecified censoring model, a second method, a doubly robust variation of outcome weighted learning, is formulated. In both instances, the treatment rule is consistent for the optimal rule when the model for either the survival times or censoring times is correctly specified. Note: it is not required that both models be correctly specified. A convex relaxation idea from support vector machines is invoked for construction of the necessary estimation algorithm.

The methodological techniques available for estimating ITRs in the single stage scenario encompass a broad spectrum. Some of these techniques have been extended to apply to the estimation of DTRs but not all have, making this an important area of future work. Because of the nature of sequential decision making, some of the associated techniques cannot easily be extended beyond the single stage setting, so it is important to continue making progress in both areas.

2.6 Analysis Techniques: Multiple Stages

There is a lot of interest and value in the practicality of creating techniques which accurately estimate optimal DTRs. The multiple stage scenario most similarly mimics the natural course of a chronic disease. Considering patients often need multiple treatments, individuals respond differently to different treatments at different points in their progression, and the longevity of the disease can be unknown, these techniques are important for an adequate treatment plan.

In order to properly define DTRs in this setting, notation will be presented that expands on that which is used in the single stage paradigm. Consider a trial with

T decision points. For $t = 1, \ldots, T$, let $A_t \in \{-1, 1\}$ be the dichotomous treatment assignment at the t^{th} stage and X_t be the patients' prognostic variables before the t^{th} decision point but after the A_{t-1} treatment assignment. The outcome, or reward, at the t^{th} stage is Y_t where larger values are assumed more desirable. Y_t is assumed to depend on all previous prognostic information $(X_1, \ldots, X_t)$, all treatment history $(A_1, \ldots, A_t)$ and previous outcomes $(Y_1, \ldots, Y_{t-1})$. The overall outcome of interest is the total reward $\sum_{t=0}^{T} Y_t$. The DTR is then a set of sequential decision rules $D = (D_1, \ldots, D_t)$ which maps from total patient history, $H_t = (X_1, A_1, \ldots, A_{t-1}, X_t)$ to the treatment space. The value function is then defined as

$$V(D) = E^D \left[\sum_{t=1}^{T} Y_t \right]$$

where E^D is the expectation under the measure P^D which is the distribution for

$(X_1, A_1, Y_1, \ldots, X_T, A_T, Y_T, X_{T+1})$ for some DTR D. The value function is the expected long term benefit if the population were to follow regimen D and can also be defined as

$$V(D) = E \left[\frac{\left(\sum_{t=1}^{T} Y_t \right) \prod_{t=1}^{T} I \left\{ A_t = D_t(H_t) \right\}}{\prod_{t=1}^{T} \pi_t(A_t, H_t)} \right].$$

Similar to the single stage situation, the value that maximizes the value function $V(D)$

$$D^* \in argmax_D V(D)$$

is the optimal DTR D^* (Zhao et al. 2014).

One of the most popular indirect methods of estimation is a computer science

method called Q-learning (Watkins and Dayan 1992). Q-learning is a form of reinforcement learning and is a dynamic programming procedure that uses backwards recursion to solve the complex Bellman equation more efficiently using regression models. In the two stage setting, Q-functions are defined as

$$Q_2(h_2, a_2) = E\left[Y|H_2 = h_2, A_2 = a_2\right], \quad Q_1(h_1, a_1) = E\left[\max_{a_2} Q_2(h_2, a_2)|H_1 = h_1, A_1 = a_1\right].$$

The Q-functions are conditional expectations where $Q_2(h_2, a_2)$ evaluates the quality of choosing treatment a_2 for patients with history h_2 and $Q_1(h_1, a_1)$ evaluates the quality of choosing treatment a_1 for patients with history h_1 assuming that the best second stage intervention has chosen. This can be extended to more than 2 phases such that

$$Q_t(h_t, a_t) = E\left[\max_{a_t} Q_{t+1}(h_{t+1}, a_{t+1})|H_t = h_t, A_t = a_t\right]$$

would evaluate the quality of choosing a_t for patients with history h_t assuming the best intervention is chosen at all future stages. In practice, these Q-functions are not known but are estimated using a linear form

$$Q_t(h_t, a_t) = h_{t,1}^T \beta_{t,1} + a_t h_{t,2}^T \beta_{t,2}.$$

In the two stage scenario, estimating the Q-functions, $\hat{Q}_t(h_t, a_t)$, is a three step procedure. First, using ordinary least squares regression, the estimates $\hat{\beta}_{2,1}$ and $\hat{\beta}_{2,2}$ are obtained by regressing the patient history on Y_2. Those estimates are used to estimate the fitted Q function at the second stage $\hat{Q}_2(h_2, a_2) = h_{2,1}^T \hat{\beta}_{2,1} + a_2 h_{2,2}^T \hat{\beta}_{2,2}$. The stage 1

pseudo outcome is $\tilde{Y}_1 = Y_1 + \max_{a_2} \hat{Q}_2(h_2, a_2)$. Note that if the outcome is only collected at the final stage (in other words, there is only one Y so there is no Y_2, Y_1), the stage 2 outcome is Y and the stage 1 psuedo outcome is $\tilde{Y} = \max_{a_2} \tilde{Q}_2(h_2, a_2)$. The first stage patient history is regressed on $\tilde{Y}_1$ to obtain the estimates $\hat{\beta}_{1,1}$ and $\hat{\beta}_{1,2}$. The first stage fitted Q function is $\hat{Q}_1(h_1, a_1) = h_{1,1}^T \hat{\beta}_{1,1} + a_1 h_{1,2}^T \hat{\beta}_{1,2}$. Finally, the estimated optimal treatment decision is given by

$$D_t^*(h_t) = argmax_{a_t} \hat{Q}_t(h_t, a_t).$$

Estimating the Q-functions is similar for three or more stage implementation where the predicted future outcome is used to create the estimates for the previous stage estimated Q-function.

While Q-learning is a very popular estimation technique, it is not without its limitations and suffers from some undesirable properties such as nonregularity, non-smoothness and asymptotic bias (Robins 2004). A nonregularity problem occurs in Q-learning when the last stage treatment is non-unique for some subjects in the population. This causes bias and inaccurate inference, causing the Q-functions to take non-linear forms. To remedy this, Goldberg et al. (2013) uses special adaptive weights within the penalization. This corrects for the non-regularity condition by concentrating on the indifference hyperplane of patient covariates where two treatment have the same effect. The indifference hyperplane is the covariate region where there is no difference between treatments. Solving this non-regularity condition involves correctly identifying the covariate values which lie on this hyperplane. This adaptive penalized Q-learning procedure can handle continuous covariates and performs better than the penalized Q-learning method. Unfortunately, it can only be implemented with discrete covariates.

Instead of the typical first stage of Q-learning (which involves solving the minimization problem) the adaptive minimization problem involves solving

$$\Phi_{2n}(\theta_2) = \sum_{i=1}^{n} \{Y_{2i} - Q_2(h_{2i}, a_{2i}; \beta_{2,1}, \beta_{2,2})\}^2 - \frac{\lambda_n}{n} \sum_{i=1}^{n} \hat{\omega}_{ni} |\beta'_{2,2} H_{2i(2)}|$$

where $\hat{\omega}_{ni}$ are the data driven weights and λ_n is the tuning parameter. Then, $\tilde{\theta}_2$ (in traditional Q learning this is the set of parameters which minimizes the ordinary least squares regression function at the second treatment decision time point) is the minimizer of Φ_{2n} and the remaining steps of Q learning are the same after substituting in $\tilde{\theta}_2$ for the normal estimator. In order to obtain the oracle property (which means the estimator behaves asymptotically as if the indifference place is already known) the selection of weights is critical. The goal is to find weights that penalize the observations that are close to or are on the indifference hyperplane and that provide weights that go asymptotically to zero for observations far from the hyperplane. This will help define where the indifference hyperplane is and resolve the non-regularity problem.

As in most statistical research areas, after developing an estimator the next step is to assess its properties, oftentimes with the use of inference techniques such as confidence intervals. When estimating optimal DTR, common approaches such as Q-learning involve estimation and interference of parameters that are non-smooth functions of the underlying generative distribution. As was mentioned before, these estimates are non-regular and asymptotically biased. Standard asymptotic approximations to the sampling distributions cannot be used to directly form reliable confidence intervals or carry out hypothesis testing (Laber et al. 2014c). One method to construct confidence intervals is an m-out-of-n bootstrap procedure to correct the nonsmoothness. The confidence sets are constructed in a way to adapt to the nonregularity

present in the underlying generative model. The data driven adaptive choice of m produces asymptotically correct confidence sets under fixed alternatives. This method has the added benefit of conceptual and computational simplicity with a corresponding R package (Chakraborty et al. 2013).

The proposed adaptive scheme to select m is a class of resample sizes given by $m = n^{f(p)}$. The suggested simple form is proposed to be

$$\hat{m} = n^{\frac{1+\alpha(1-\hat{p})}{1+\alpha}}$$

where $\alpha > 0$ is a tuning parameter. α controls the smallest acceptable sample size and may be dictated by practical considerations or tuned using the data. A bootstrap algorithm is used for choosing α using data which appears to reduce conservatism. When the parameter of interest is a linear function of the parameters, $(c'\theta_{1,n})$, the algorithm first draws B_1 m-out-of-n first stage bootstrap samples and estimates $c^T\hat{\theta}^{b_1}_{1,n}$. α is fixed at the smallest value in the grid. $\hat{m}^{b_1}$ is then calculated using the equation above. Repeat this drawing B_2 $\hat{m}^{b_1}$-out-of-n second stage bootstrap samples and calculate $c^T\hat{\theta}^{(b_1,b_2)}$ which is a double bootstrapped version of the estimate. For all b_1, compute $\left(\frac{\eta}{2}\right)$x100 and $\left(1-\frac{\eta}{2}\right)$x100 percentiles which are the lower bounds and upper bounds defined as $\hat{l}^{b_1}_{DB}$ and $\hat{u}^{b_1}_{DB}$ respectively. The coverage rate of the double bootstrap confidence interval from all first stage bootstrap data sets is

$$\frac{1}{B_1}\sum_{b_1=1}^{B_1} I\left(c^T\hat{\theta}^{b_1}_{1,n} - \frac{\hat{u}^{b_1}_{DB}}{\sqrt{\hat{m}^{b_1}}} \leq c^T\theta_{1,n} \leq c^T\hat{\theta}^{b_1}_{1,n} - \frac{\hat{l}^{b_1}_{DB}}{\sqrt{\hat{m}^{b_1}}}\right).$$

Increase α to the next highest value on the grid until the coverage rate is at or exceeds the nominal value and in that case pick the current value of α as the final value. The process is repeated until the coverage rate of the double bootstrap confidence interval

attains a nominal coverage rate or all the options on the grid are exhausted.

Another methodology that can accommodate the non-regularity from using Q-learning to estimate parameters is the locally consistent Adaptive Confidence Interval (ACI) (Laber et al. 2014c). When construction of DTRs using Q-learning, there is particular interest in reducing bias of first stage coefficients. If the Q-function is near 0 with high probability there will be issues approximating the distribution of $\sqrt{n}\left(\hat{\beta}_1 - \beta_1^*\right)$. Once the asymptotically biased parameters are identified, given the correct amount of shrinkage, a shrinkage estimator can reduce the bias. However, shrinking too aggressively leads to bad performance in finite samples. Constructing valid confidence intervals for non-regular estimators is a difficult task because estimating the sampling distribution of the estimator cannot be done uniformly. The proposed solution is a locally consistent confidence interval for linear combinations of the first stage coefficients. The interest is not in construction of second stage confidence intervals because they can be estimated using standard methods for least square estimators. Since it is not possible to construct a uniformly convergent estimator of the limiting distribution of $\sqrt{n}\left(\hat{\beta}_1 - \beta_1^*\right)$, for a given constant c the proposed method bounds $c^T\sqrt{n}\left(\hat{\beta}_1 - \beta_1^*\right)$ between two regular uniformly convergent upper and lower bounds. These smooth bounds can be bootstrapped to form a confidence set for $c^T\beta_1^*$. The extension to more than two stages is straightforward as the last stage uses standard methods for least squares estimation, so the ACI would be used on all previous stages.

Similar to previously discussed analysis techniques, for medical research it is important to develop these techniques to accommodate censored data. Goldberg and Kosorok (2012) developed a Q-learning algorithm that allows for censored data when the outcome of interest is survival time and allows for a flexible number of stages in a randomized trial. Q-learning is expanded upon by using inverse probability censoring weighting to account for censored observations.

For each $t = 1, ..., T$, let the state S_t be the pair $S_t = (X_t, Y_{t-1})$ where X_t is either a vector of covariates describing the condition of the patient before time t or it is null. If X_t is null then a failure happened during the t^{th} stage. Let Y_{t-1} be the length of time between decision points t and $t-1$. Hence, $\sum_{j=1}^{t} Y_j$ is the total survival time, or reward, up to and including stage t. Let $C \in [0, \tau]$ be the censoring variable. The goal is to find a policy that maximizes the expected rewards. Then, the optimal policy, D^*, is the one that approximately maximizes over all policies of $E_{0,\pi}\left[(\sum_{t=1}^{T} Y_t) \wedge \tau\right]$ where $\bar{T}$ is the random number of stage for the subject. This optimal policy is found using a three step algorithm. First the problem is mapped to an auxiliary problem. The auxiliary problem creates modified trajectories of a fixed length T and the modified sum of the rewards is less than or equal to τ to account for censoring. Next, the Q-functions are approximated $\{\hat{Q}_1, ..., \hat{Q}_T\}$ using the original Q-function framework. Last, the optimal treatment rule, D^*, is found by maximizing $\hat{Q}_t$ over all possible a_t.

Recall the methodology introduced in the previous section for estimation of ITRs in a single stage. Two of those methods will be expanded on when estimating DTRs for multiple stages of treatment. In the single decision scenario presented by Zhang et al. (2012b) the estimation procedure was restructured into a classification problem. In this case the optimal classifier corresponds to the optimal treatment decision. The optimal classifier was found by estimating the Bayes classifier which is the one that minimizes the expected weighted misclassification error. This can be expanded upon for the two decision point scenario based on reassessing the problem as a monotone coarsening problem using an augmented inverse probability weighted estimator (AIPWE) to estimate the mean outcome (Zhang et al. 2013). Assign Y^* to be the often unobserved potential outcome and Y_D^* to be the potential outcome associated with treatment regime D. The optimal treatment regime D^* is that which satisfies $E[Y^*(D^*)] \geq E[Y^*(D)]$ meaning it is that which maximizes the expected potential outcome. The problem is cast into a

monotone coarsening problem where the coarsening happens at random if, for each t, the probability that the data are coarsened at level t given the full data depends only on the data observed at level t. Then, from Robins et al. (1994), under these coarsening assumptions if the coarsening mechanism is correctly defined then asymptotically linear consistent estimators for $E[Y^*(D_\eta)]$ for a fixed η have the form

$$\frac{\sum_{i=1}^n I(C_{\eta,i} = \infty)}{K_{\eta,k}\bar{X}_{k,i}} Y_i + \frac{\sum_{i=1}^n \left\{ I(C_{\eta,i} = k) - \lambda_{\eta,k}(\bar{X}_{k,i}) I(C_{\eta,i} > k) \right\}}{K_{\eta,k}(\bar{X}_{k,i})} L_k(\bar{X}_{k,i})$$

where $L_k(\bar{X}_{k,i})$ are arbitrary functions, $C_{\eta,i}$ is the discrete coarsening variable, $K_{\eta,K}(\bar{X}_K) = \prod_{k'=1}^k \left\{1 - \lambda_{\eta,k'}(\bar{X}_{k'})\right\}$ and $\lambda_{\eta,k'}(X'_k)$ is the hazard function. The left side of the above estimator is on its own a consistent estimator if $\lambda_{\eta,k}(\bar{X}_k)$ is correctly specified. Then the entire estimator is a doubly consistent robust estimator for $E[Y^*(D\eta)]$ if either $\lambda_{\eta,k}(\bar{X}_k)$ are correctly specified or if $L_k(\bar{X}_{k,i}) = Y^*(D_\eta) | \left\{ \bar{X}_k^*(\bar{D}_{\eta_{k-1}}) = \bar{x}_k \right\}$.

O-learning was presented as a machine learning approach which directly optimizes the value function $V(D)$ where each subject's weight is proportional to their clinical outcome. This is a weighted classification error problem since finding the D^* that maximizes $V(D)$ is equivalent to finding the D^* that minimizes $\bar{V}(D)$. O-learning can also be expanded to the two stage paradigm using a few strategies. One such way is backwards outcome weighted learning (BOWL) which modifies existing algorithms to solve a sequence of weighted classification problems (Zhao et al. 2014). The algorithm is backwards fitting and at each time point T, the algorithm is as follows. The goal in the first stage is to minimize

$$\frac{n^{-1} \sum_{i=1}^n \left[Y_{iT} \phi \left\{ A_{iT} f_T(H_{iT}) \right\} \right]}{\pi_T(A_{it}, H_{it})} + \lambda_{T,n} \|f_T\|^2$$

with respect to f_T where $\hat{f}_T$ is the minimizer. The optimal decision rule is

$$\hat{D}_T(h_T) = sign\left\{\hat{f}_T(h_T)\right\},$$

and this stage is essentially equivalent to the single stage outcome weighted learning found in Zhao et al. (2012) and has a similar dual objective function as found in support vector machines. The second stage is, for $t = T-1, T-2, \ldots, 1$, to backward sequentially minimize

$$n^{-1}\sum_{i=1}^{n}\frac{\left(\sum_{j=1}^{T}Y_{ij}\right)\prod_{j=t+1}^{T}I\left\{A_{ij}=\hat{D}_j(H_{ij})\right\}}{\prod_{j=1}^{T}\pi_j(A_{ij},H_{ij})}x\phi\left\{A_{ij},f_t(H_{it})\right\}+\lambda_{t,n}\|f_t\|^2$$

where $\hat{D}_{t+1},\ldots,\hat{D}_T$ are previously obtained.

A disadvantage of BOWL is the number of observations utilized by the algorithm decreases geometrically as t decreases. The authors explain this can be solved using iterative outcome weighted learning (IOWL) which involves re-estimating the optimal treatment rule at stage 2 after the stage 1 rule is estimated. This estimate is based on the subset of patients whose stage 1 treatment assignments are consistent with the optimal rule. The procedure would continue with a re-estimation of the stage 1 treatment rule based on the new optimal stage 2 rule. IOWL allows the exploration of different subjects through iterative re-estimation.

Zhao et al. (2014) also present simultaneous outcome weighted learning (SOWL) which frames estimation of DTRs as a single classification problem. This is an effective way of looking at the problem because a multiple stage treatment plan has not previously been estimated simultaneously using a single algorithm. The method directly

optimizes the empirical counterpart of the value function in one step. Since this problem is computationally difficult (mostly because of the discontinuity of the indicator functions) a continuous and concave surrogate function is used in lieu of the product of indicators that would usually be required. In the two decision point scenario, the surrogate reward function is chosen to mimic hinge loss: $\psi(Z_1, Z_2) = min(Z_1 - 1, Z_2 - 1, 0) + 1$ where $Z_1 = A_1 f_1(H_1)$ and $Z_2 = A_2 f_2(H_2)$. Hence the SOWL estimator maximizes

$$n^{-1} \sum_{i=1}^{n} \frac{(\sum_{j=1}^{2} Y_{ij}) \psi \left\{ A_{i1} f_1(H_{i1}), A_{i2} f_2(H_{i2}) \right\}}{\prod_{j=1}^{2} \pi_j(A_{ij}, H_{ij})} - \lambda_n \Big(\|f_1\|^2 + \left\|f_2\right\|^2 \Big),$$

where the tuning parameter λ_n controls the amount of penalization. This can easily be extended to more than 2 stages.

Much exciting and significant work is being done in developing treatment rules with the intent of dynamic sequential decision making. To this effect, there have been promising advancements, but these methods often times need to be expanded upon or adapted to various specific settings. Science will forever be changing and the best everyone can hope to do is keep up. While existing methodology can always be improved and generalized, at the same time there will probably never be a lack of need for new and innovative mechanisms for estimating optimal DTRs.

2.7 Observational Data

As clinical researchers were experimenting with new trial designs to improve patient treatment plans, epidemiologists were simultaneously investigating the relationships of time varying continuous exposures to various outcomes in observational data (Lavori and Dawson 2014). They developed a longitudinal generalization of Rubin's potential outcomes (Holland 1986) for inference between exposure and outcomes for

observational data which naturally led to an interest in developing DTRs for this data. These methods include G-estimation, non-parametric theory and methods designed to handle large mathematical combinations of observational treatment regimes. These treatment plans are germane for clinical patient treatment because these DTRs assume the exposures are assigned in a way that is conditionally independent of the potential future responses given the history of the patients and treatments up to the current state. This resembles the assumptions made when developing treatment plans using randomized prospective data.

Randomized trials have been regarded as the optimal way to test any treatment, so why would it be a good idea to use observational data to develop and determine the optimal DTR instead of designing a SMART? Often times situations arise where a randomized trial is impossible or impractical, so it is efficacious to perform an observational study instead. Additionally, observational data may exist from a preexisting study. Using this resource can be more expedient and reduce significant financial burden because new patients are not needed and no treatments are given. The development of DTRs is often exploratory and hence it is potentially important to be able to estimate these treatment plans using large samples of observational data with the intention of validating the DTR in a confirmatory randomized trial. Furthermore, collecting observational data on time-varying outcomes, predictors and confounders can sometimes emulate a randomized trial that lacks baseline randomization.

An effective estimation technique when eliciting DTRs from observation data is the parametric g-formula. The parametric g-formula uses Robins' G-estimation to naturally estimate DTRs and can appropriately adjust for time-dependent cofounding variables (Young et al. 2011). This formula is an alternative to inverse probability weighting that provides more efficient estimates but requires more parametric modelling assumptions. When using low dimensional data, the g-formula is

$$\sum_{l_k}\sum_{j=0}^{k} P(Y_{j+1}=1|L_j=l_j, A_j=a_j^D, \bar{Y}_j=C_{j+1}=0)$$
$$\times \prod_{s=0}^{j}[P(Y_s=0|L_{s-1}=l_{s-1}, A_{s-1}=a_{s-1}, Y_{s-1}=C_s=0)$$
$$\times f(l_s|l_{s-1}, a_{s-1}, Y_s=C_s=0)]$$

where Y_j represents the indicator of death by the end of time j, L_j represents the last measurements of covariates preceding treatment assignment, A_j indicates treatment before time j, and C_j indicates censorship status by the end of time j such that $P(Y_s=0|L_{s-1}=l_{s-1}, A_{s-1}=a_{s-1}, Y_{s-1}=C_s=0)$ is the probability of surviving through month s conditional on not being censored through s, surviving through $s-1$ and adhering to the designated treatment regime through $j-1$ for the specific history (L_j, A_j), $f(l_s|l_{s-1}, a_{s-1}, Y_s=C_s=0)$ is the density for L_s conditional on not being censored and survival through s and adhering to the designated treatment regime through $j-1$ for the specified history. It can be computed for the potential outcome by non-parametrically estimating the value of each density function for all of the possible histories of patients. The formula takes the sum over the histories but requires that all possible covariates need to be categorical. For high dimensional data, the g-formula can only be carried out by estimating the density functions using parametric modelling assumptions then taking the sum over the histories via Monte Carlo simulation. Because of distributional a priori knowledge for certain histories, when estimating the g-formula parametric models are not needed to be imposed over all components of the densities and histories.

While there has been progress using g-estimation for estimating optimal DTRs, there exists limited methods to check the model performance and validate assumptions.

Fortunately, Rich et al. (2010) developed a method for g-estimation model checking diagnostics that uses traditional tools when evaluating DTRs. With the goal of assessing the model examining a residual plot, a new residual is proposed that redefines a fitted $\hat{Y}_i$ (since the residual is defined as $Y_i - \hat{Y}_i$). The authors propose

$$\hat{Y}_{ij}(\psi) = E\left[H_{ij}|\bar{L}_j, \bar{A}_{j-1}; \xi(\psi_j)\right] - \sum_{m=j+1}^{K-1} \mu_m(\psi_m) + \gamma_j(\psi_j)$$

where ψ is the blip parameter, L is the history, H is interpreted as the difference between removing the effect of the treatment from the outcome and adding the effect of making the optimal treatment decisions in the future, μ is the effect of the optimal treatments in the future, γ is the blip function, and ξ is a nuisance parameter. The blip function is a functional form of the mean difference in responses under two possible actions conditional on the patients' history. The estimates of $\hat{\psi}$ and $\hat{\xi}_j(\hat{\psi}_j)$ from the G-estimation procedure can then be plugged in to obtain a usable fitted outcome. Then, the residual can be as usual: $r_{ij} = Y_i - \hat{Y}_{ij}(\hat{\psi})$. These residuals can be used to diagnose underlying specification problems in the blip and expected potential outcome models and check linearity assumptions. The residual plots can be visually assessed as usual where a good model will have a symmetric distribution around zero and no trend when plotted against covariates or fitted values.

G-estimation in the context of estimating DTRs has advantages over traditional parametric approaches for producing consistent estimators. However, these estimators are asymptotically biased under a given structural nested mean model for certain data distributions (coined exceptional laws) and exhibit non-regular behavior. To combat this, Moodie and Richardson (2009) presented a new approach called Zeroing Instead of Plugging In (ZIPI). ZIPI provides estimators nearly identical to those provided by g-estimation but with the benefit of reducing bias in those situations when decision

rule parameters are not shared across intervals. More specifically in the context of constructing DTRs, the observed longitudinal distribution function is "exceptional" if at some interval there is a positive probability that the true optimal decision rule is not unique. For a distribution to be exceptional, the blip function must include at least one covariate (such as the previous treatment), and the probability that the true blip function has value 0 is positive. The proposed ZIPI method is considered a modification of g-estimation when there is no parameter sharing and detects and reduces bias in the presence of exceptional laws. Bias is found in the g-estimating equation of ψ_1 by including the upwardly biased estimate of

$$I\left\{g_2(\bar{L}_2, A_1; \psi_2) > 0\right\} g_2(\bar{L}_2, A_1; \psi_2)$$

into the g-estimating equation when $g_2(\bar{L}_2, A_1; \psi_2)$ is close to 0. The proposed algorithm searches for individuals who will likely have $g_2(\bar{L}_2, A_1; \psi_2) = 0$ and then uses the best guess of zero instead of the estimate obtained by plugging in ψ_2. Hence, it uses an all or nothing weight system which either applies weights 0 or 1 depending on the estimated unique rule class membership. This method is considered a class of pre-test estimators frequently used in statistical analysis. ZIPI requires testing all the individuals at all intervals except the first ones which creates a concern about potentially reduced power.

What about modifying preexisting techniques developed for randomized data to be used on observational data? Instead of developing an array of new techniques, is it possible to alter portions of those methods so they are applicable for observational data? Unfortunately, the answer to that question is unclear as this has not been a well-studied area. However, Moodie et al. (2012) extended one of the more frequently utilized methods of optimal DTR estimation, Q-learning, to accommodate observational data. A soft threshold approach is used which has a good performance in terms of bias

and coverage in the non-regular settings. This approach shrinks the problematic term in the potential outcome towards zero. In the Q-learning algorithm this would replace the potential outcome with

$$\hat{Y}_{1,i}^{ST} = \beta_2^T H_{20,i} + |\hat{\Psi}_2 H_{21,i}| \cdot \left\{ 1 - \left(\frac{\lambda_i}{|\hat{\Psi}_2^T H_{21,i}|^2} \right)^+ \right\},$$

where λ is a data driven tuning parameter. λ_i is chosen using a Bayesian approach where $\lambda_i = 3H_{21,i}^T \Sigma_2 H_{21,i}/n$ and Σ_2 is the estimated covariance matrix of $\hat{\Psi}_2$. The rest of the Q-learning algorithm stays the same.

When using Q-learning for observational data, the basic approach requires the construction of a propensity score, $\pi(x) = P(A = 1|X = x)$, or treatment model followed by some form of adjustment. It assumes the treatment received is independent of known covariates given the propensity score. This leads to unbiased estimates of the treatment effect based on the conditional expectation modelling the outcome given the propensity score. In inverse probability weighting analysis, the weights are used to create a pseudo-sample so that the treatment does not depend on the variables in the pseudo-sample. Including covariates into the models for the Q-functions can be implemented in four ways which perform well: including the covariates as linear terms in the Q-function, including the propensity score as a linear term in the Q-function, including quintiles of the interval-specific propensity score (which depends on a time varying confounding variable) as covariates in the j^{th} interval Q-function, and IPTW weighted with H_1 and H_2 defined as in traditional Q-learning.

Not all data can be collected using a randomized clinical trial, so it is imperative to develop methodologies that can estimate optimal DTRs using this observational data. This data is difficult because there is no randomization, but the assumption of

independence between exposure and predicted future outcome simplifies the problem to an extent. While the parametric g-formula and Q-learning are great options for analysis, not all data will benefit from these techniques and more work would provide more resources and flexibility for scientists.

2.8 Competing Outcomes

While development of these advanced estimation techniques is necessary to effectively personalize medical care, treatment of the entire person, not just one disease, should be considered as well. In the tangible clinical setting, the patient or caregiver will likely be interested in balancing competing outcomes such as survival, quality of life and financial burden. While survival may be the ultimate goal for a cancer patient, a single mother of two may prefer a treatment that allows her to work (higher quality of life) which could lead to a longer treatment course, or patients may need to balance the financial burden with their treatment plan. This is a very new area of study inside the personalized medicine umbrella, but it is important and quickly developing.

A crucial step in balancing competing outcomes for personalized patient care is eliciting the patient's or physician's preferences regarding the ideal tradeoff between the outcomes. Lizotte et al. (2012b) produced an inverse preference elicitation approach which first considers all of the actions available at any given state. Then, for each action, asks what range of preferences makes that action a good choice. This provides a large amount of information about the potential actions at each state. The patients also have the ability to see if their preference is near the boundary or see if a small change in preference results in a change of recommended treatment. In this situation, the patient can feel confident that both treatments perform well and make the decision based on other potentially minor preferences. This method provides an efficient algorithm that

computes the optimal policy for varying reward functions and provides insight into how the choice of reward influences the optimal treatment decision.

Let $V_t(h_t, \delta_t)$ be the value function where δ_t represents the patient's preference that can be thought of as the part of the patient's history that doesn't evolve with time and does not influence transition dynamics between time points. The fixed δ identifies each single decision process by fixing a reward function which can be viewed as representing the patient's preference. $V_t(h_t, \delta_t)$ is a piecewise linear function and is developed through a series of state t Q-functions obtained from the Bellman equation. The exact piecewise linear representation of the value function for each history and time point is found allowing efficient, exact computation of value backups for all δ_t. This method also identifies the actions not optimal for any history or preference. These value backups require minimization over all possible actions and expectations over all future states. Convex hull routines provide ordered output which makes it easy to recover the list of treatments that are optimal for each δ_t and what values of δ_t correspond to a change in the optimal treatment strategy. These values of δ_t are the knots in the piecewise linear representation and define the piecewise function. A list of knots with their corresponding values is used instead of the list of points. The policy is found in a list which contains the optimal corresponding actions at each knot and the optimal policy for each stage by taking the intersection of the treatment lists for the endpoints of the segment. A binary search for the largest knot that is less than some δ_t defines which linear piece is maximized for δ_t and hence only that single linear function needs to be evaluated. The piecewise linear representation of the value function $V_t(h_t, \delta_t)$ is used to efficiently compute a piecewise linear approximation of the Bellmen equation by evaluation conditional expectations of $V_t(h_t, \delta_t)$ over possible future states. This can easily be generalized to a scenario where there are an arbitrary number of features of state variables and where there are more than two decision points.

Alternatively, Laber et al. (2014b) developed a way to construct DTRs that does not require tradeoffs between outcomes by eliciting a clinically significant difference for each respective outcome. When the algorithm concludes that no single treatment is best, the patient's or doctor's preferences are able to be incorporated. They are free to choose the treatment arbitrarily based on other qualities that matter to them, such as cost. This method involves set-valued dynamic treatment regimes that take as input the current patient history and provide as output a set of recommended treatments.

Considering just the static set valued decision rules for the single decision time point, let Y and Z be the competing outcomes and Δ_Y, Δ_Z represent the predetermined clinically meaningful difference in the respective outcomes. In the ideal situation, the algorithm will produce one recommended treatment if that treatment provides significant benefit to one outcome without producing significant detriment to the other. However, in all other cases, the algorithm will produce a set of recommended treatments and the decision is left up to the clinician or patient. With

$$\tau_Y(h) = E\left[Y|H = h, A = 1\right] - E\left[Y|H = h, A = -1\right]$$

and

$$\tau_Z(h) = E\left[Z|H = h, A = 1\right] - E\left[Z|H = h, A = -1\right]$$

then the ideal decision rule $\pi_\Delta^{ideal}(h)$ is either

1. $sign\{\tau_Y(h)\}$ if $|\tau_Y(h)| > \Delta_Y$ and $sign\{\tau_Y(h)\}\tau_Z(h) > -\Delta_Z$
2. $sign\{\tau_Z(h)\}$ if $|\tau_Z(h)| > \Delta_Z$ and $sign\{\tau_Z(h)\}\tau_Y(h) > -\Delta_Y$
3. $\{-1, 1\}$ otherwise

Generalizing this procedure to dynamic set valued decision rules for two or more decision points, the algorithm is backwards regressive and Q-learning with linear working models is used to estimate $r_Y(h)$ and $r_Z(h)$. Using ordinary least squares, the optimal treatment set can be estimated from patient history. At the second stage, estimating $\pi_{2\Delta}^{ideal}$ is essentially the same as described for the single decision point situation. To find $\pi_{2\Delta}^{ideal}$, it is assumed that the best single treatment decision (not a set-valued decision) was made, τ_2. Then, $\pi_{1\Delta}^{ideal}(h_1, \tau_2)$, at the first stage is

1. $sign\{\tau_Y(h_1, \tau_2)\}$ if $|\tau_Y(h_1, \tau_2)| > \Delta_Y$ and $sign\{\tau_Y(h_1, \tau_2)\}\tau_Z(h_1, \tau_2) > -\Delta_Z$
2. $sign\{\tau_Z(h_1, \tau_2)\}$ if $|\tau_Z(h_1, \tau_2)| > \Delta_Z$ and $sign\{\tau_Z(h_1, \tau_2)\}\tau_Y(h_1, \tau_2) > -\Delta_Y$
3. $\{-1, 1\}$ otherwise

Hence, the optimal decision rule is the set valued rule

$$\pi_{1\Delta}^{ideal}(h_1) = \bigcup_{\tau_2 \epsilon C\left(\pi_{2\Delta}^{ideal}\right)} \pi_{1\Delta}^{ideal}(h_1, \tau_2)$$

where $C\left(\pi_{2\Delta}^{ideal}\right)$ is the set of all treatment options compatible with $\pi_{2\Delta}^{ideal}$ and τ_2 is compatible with π_2 if and only if $\tau_2(h_2) \epsilon \pi_2(h_2) \quad \forall \quad h_2$.

Progressing from estimating techniques for one outcome to multiple outcomes is the next step in the natural course of this specialization. Practically, patients and doctors will have more than one goal when developing a treatment plan and these complicated preferences should be taken into consideration if possible. As this is a new area, there is continuing progress.

2.9 Conclusion

Data collected from SMARTs is an integral part of effectively developing DTRs. This has been a hot topic in clinical trial design in recent years and appears to be the future of clinical practice. They are flexible, informative studies which utilize all of the participants and can answer more questions than a traditional RCT. Pilot studies are essential for designing a SMART so that resources are optimized and the essential information is properly collected. While there has been progress made in properly designing SMARTs, there is still a lot of work to be done particularly in sample size estimation and handling missing data. A plethora of direct and indirect techniques have been presented here to highlight some of the most current methods available so the reader can make informed decisions when creating their analysis plan for a SMART design or even when working with observational data. One of many future directions in this area is practical implementation in clinical practice which involves estimating DTRs for competing outcomes, an area which is quickly expanding. The recent progress over the last ten years has been very exciting, but there are still many areas that could use further development and many topics that have not been explored yet. The future of implementing SMARTs for developing DTRs to personalize medicine is bright and promising.

CHAPTER 3: INCORPORATING PATIENT PREFERENCES INTO ESTIMATION OF OPTIMAL INDIVIDUALIZED TREATMENT RULES

3.1 Introduction

It is widely recognized that best possible clinical care is tailored to individual patient characteristics (Sox et al. 2008, Hamburg and Collins 2010, Collins and Varmus 2015) including subjective factors like patient personal preference (Edwards and Elwyn 2009). Individualized treatment rules (ITRs) formalize personalized clinical care as a function from up-to-date patient information to a recommended treatment. An optimal ITR maximizes the mean of some pre-specified clinical outcome if applied to make treatment decisions for all patients in a population of interest. This definition of optimality is mathematically convenient as it reduces estimation of an optimal ITR to a scalar optimization problem over a class of potential ITRs. However, this formulation does not directly allow for shared decision making wherein patient preferences are integrated into the decision process (Barry and Edgman-Levitan 2012, Drake et al. 2010); on the other hand, direct preference elicitation in which the patient chooses parameters indexing a composite outcome is not feasible unless patients have undergone specialized training (Brennan 1998, Braziunas 2006, Lizotte et al. 2012a). Thus, a common approach for preference elicitation is to administer a questionnaire populated with items that are accessible (meaningful) to a patient in a domain context yet are informative about a preferences in the outcome space. These questions may ask the patient to rank or numerically score different health states and/or may ask about their attitudes or

experiences with certain outcomes. We consider the simplest possible setting in which the questionnaire comprises a series of binary questions; we use item response theory (Embretson and Reise 2013) to estimate a conditional distribution over preferences given a patient's answers to the items in the questionnaire. We use this conditional distribution to derive a preference-sensitive optimal ITR for each patient.

There is a vast literature on estimation for optimal ITRs assuming a single scalar outcome. Estimators are typically broadly categorized as regression-based (Murphy 2005b, Henderson et al. 2010, Zhao et al. 2011, Qian and Murphy 2011, Goldberg and Kosorok 2012, Moodie et al. 2012; 2014, Tian et al. 2014, Laber et al. 2014a) or policy-search-based (Orellana et al. 2010, Zhang et al. 2012a, Zhao et al. 2012, Zhang et al. 2012b; 2013, Zhao et al. 2014; 2015, Laber and Zhao 2015), though this division is somewhat superficial (Taylor et al. 2015). The estimators referenced above assume that a single outcome has be pre-specified and that this outcome faithfully represents the preferences of all patients in the population of interest.

The literature that does not assume a fixed and known composite outcome is scarce and none addresses patient-specific preferences directly through elicitation. There are three primary approaches to estimation of optimal ITRs with multiple outcomes: (i) set-valued treatment regimes; (ii) inverse-preference elicitation; and (iii) constrained estimation. Set-valued treatment regimes (Laber et al. 2014b, Lizotte and Laber 2016) map current patient information to a subset of possible treatment options that are not estimated to be uniformly worse across all outcomes; the expectation is that a single treatment will be selected from the recommended set using patient preferences and clinical judgment (no guidance is provided on how this will be done). Thus, set-valued treatment regimes incorporate patient preferences through selection from the recommended subset but do not directly elicit preferences. Inverse-preference elicitation (Lizotte et al. 2012a) attempts to communicate to a patient the implicit composite

outcomes that would be consistent with each possible treatment choice, e.g., choosing treatment 1 over treatment -1 corresponds to valuing side-effect burden at least three times as much as efficacy. Application of inverse preference elicitation requires the patient to interpret convex combinations of outcomes which may be difficult without specialized training. Another approach to estimation of optimal ITRs with multiple outcomes is to maximize the expectation of one outcome subject to constraints on a functional of the distribution of the the remaining outcomes (Linn et al. 2016). This approach assumes that a single (unknown) composite outcome reflects patient preferences across the entire population.

The proposed methodology incorporates patient preferences into treatment selection in a way that is mathematically rigorous yet does not require the patient to undergo specialized training or understand complex quantitative concepts. This has at least two important clinical impacts: (i) it facilitates 'patient-centered care' in which patients play a key role is decision making and the evaluation of their own health outcomes; and (ii) it offers a principled means for matching patient preferences to an optimal treatment based on potentially complex outcome profiles. While many clinical and intervention scientists already seek to implement patient-centered care and to personalize treatment decisions based on patient preferences, it is recognized that the rapid introduction of new therapies and a high-degree of patient preference heterogeneity makes this a difficult task to perform using expert judgement alone (Smoller and Nierenberg 1999, Basu and Meltzer 2007, Frank and Zeckhauser 2007, Hodgkin et al. 2012, Huskamp et al. 2013). The proposed methodology can be viewed as a tool for clinical decision support in the context of patient-centered care.

3.2 Optimal ITRs with Heterogeneous Patient Preferences

3.2.1 Setup and Notation

We assume that the observed data, $\{(\boldsymbol{W}_i, \boldsymbol{X}_i, A_i, Y_i, Z_i)\}_{i=1}^n$, comprises n independent and identically distributed tuples $(\boldsymbol{W}, \boldsymbol{X}, A, Y, Z)$, one per subject, where: $\boldsymbol{W} \in \{0,1\}^p$ denotes answers to items in a preference questionnaire; $\boldsymbol{X} \in \mathbb{R}^m$ denotes pre-treatment patient covariates; $A \in \{-1, 1\}$ denotes the assigned treatment; $Y, Z \in \mathbb{R}$ denote outcomes of interest, coded so that higher values are better. In the context of our application to schizophrenia: $\boldsymbol{W}$ denotes subject's responses to a subset of questions on the Hogan Drug Attitude Inventory (Hogan et al. 1983); $\boldsymbol{X}$ denotes patient demographics, measures of symptom severity, and the presence/absence of comorbidities (see Section 3.4 for a complete description); A denotes perphenazine or atyptical antipsychotic medication; Y denotes efficacy measured in terms of the Positive and Negative Syndromes Scale (PANSS); and Z denotes a measure of side-effect severity. In this context, an ITR, $\pi : \operatorname{dom} \boldsymbol{W} \times \operatorname{dom} \boldsymbol{X} \to \operatorname{dom} A$, is a map from answers on a completed questionnaire and patient covariates to a recommended treatment; under π a patient with responses $\boldsymbol{W} = \boldsymbol{w}$ and covariates $\boldsymbol{X} = \boldsymbol{x}$ would be recommended treatment $\pi(\boldsymbol{w}, \boldsymbol{x})$.

To define an optimal ITR, we assume that each individual in the population possesses a latent preference, denoted $E \in \mathbb{R}$, which indexes a utility function, $U(y, z; e)$, that induces a total ordering on $\operatorname{dom} Y \times \operatorname{dom} Z$ so that a patient with preference $E = e$ would prefer outcomes (y, z) to (y', z') if $U(y, z; e) \geq U(y', z'; e)$. Let $Y^*(a)$ and $Z^*(a)$ denote the potential outcomes under treatments $a \in \{-1, 1\}$ (Rubin 1978) so that $U\{Y^*(a), Z^*(a); E\}$ is the potential utility function under treatment a. For any ITR, π, define the potential utility as $V_U(\pi) = \mathbb{E}\left[\sum_{a \in \{-1,1\}} U\{Y^*(a), Z^*(a); E\} \, 1_{\pi(\boldsymbol{W}, \boldsymbol{X}) = a}\right]$.

The optimal regime, π_U^{opt}, satisfies $V_U\left(\pi_U^{\text{opt}}\right) \geq V_U\left(\pi\right)$ for all ITRs π.

We assume a utility of the form $U(y,z;e) = \Phi(e)y + \{1-\Phi(e)\}z$ where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal random variable. The next result shows that this assumption incurs only a small loss of generality. Define

$$\begin{aligned}
R_Z(\boldsymbol{w},\boldsymbol{x}) &= \mathbb{E}\left\{Z^*(1)\middle|\boldsymbol{W}=\boldsymbol{w},\boldsymbol{X}=\boldsymbol{x}\right\} - \mathbb{E}\left\{Z^*(-1)\middle|\boldsymbol{W}=\boldsymbol{w},\boldsymbol{X}=\boldsymbol{x}\right\},\\
R_Y(\boldsymbol{w},\boldsymbol{x}) &= \mathbb{E}\left\{Y^*(1)\middle|\boldsymbol{W}=\boldsymbol{w},\boldsymbol{X}=\boldsymbol{x}\right\} - \mathbb{E}\left\{Y^*(-1)\middle|\boldsymbol{W}=\boldsymbol{w},\boldsymbol{X}=\boldsymbol{x}\right\},\\
R_U(\boldsymbol{w},\boldsymbol{x}) &= \mathbb{E}\left[U\left\{Y^*(1),Z^*(1);E\right\}\middle|\boldsymbol{W}=\boldsymbol{w},\boldsymbol{X}=\boldsymbol{x}\right]\\
&\quad -\mathbb{E}\left[U\left\{Y^*(-1),Z^*(-1);E\right\}\middle|\boldsymbol{W}=\boldsymbol{w},\boldsymbol{X}=\boldsymbol{x}\right],
\end{aligned}$$

we assume that the foregoing quantities are well-defined for all $\boldsymbol{w}$ and $\boldsymbol{x}$. It can be shown that $\pi_U^{\text{opt}}(\boldsymbol{w},\boldsymbol{x}) = \text{sign}\left\{R_U(\boldsymbol{w},\boldsymbol{x})\right\}$ (Qian and Murphy 2011). .

Lemma 3.2.1. *Assume that* $\max\left\{R_U(\boldsymbol{w},\boldsymbol{x})R_Z(\boldsymbol{w},\boldsymbol{x}),\, R_U(\boldsymbol{w},\boldsymbol{x})R_Y(\boldsymbol{w},\boldsymbol{x})\right\} > 0$ *for all* $\boldsymbol{x}, \boldsymbol{w}$. *Then, there exists a real-valued random variable,* E', *such that: (i)* $E' \perp A, \left\{\left\{Z^*(a),Y^*(a)\right\} : a \in \{-1,1\}\right\}\middle|\boldsymbol{X},\boldsymbol{W}$; *and (ii) the ITR*

$$\pi_{\text{CVX}}^{\text{opt}}(\boldsymbol{x},\boldsymbol{w}) = \arg\max_{a\in\{-1,1\}} \mathbb{E}\left[\Phi(E')Y^*(a) + \left\{1-\Phi(E')\right\}Z^*(a)\middle|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{W}=\boldsymbol{w}\right]$$

satisfies $V_U(\pi_{\text{CVX}}^{\text{opt}}) = V_U(\pi_U^{\text{opt}})$.

Proof. Define $\xi(\boldsymbol{w},\boldsymbol{x}) = \left|R_Z(\boldsymbol{w},\boldsymbol{x})/R_Y(\boldsymbol{w},\boldsymbol{x})\right|^{\text{sign}\{(R_U(\boldsymbol{w},\boldsymbol{x})R_Z(\boldsymbol{w},\boldsymbol{x})\}}$ if both $R_Z(\boldsymbol{w},\boldsymbol{x})$ and $R_Y(\boldsymbol{w},\boldsymbol{x})$ are nonzero, and $\xi(\boldsymbol{w},\boldsymbol{x}) = 1$ otherwise. Set $E' = E'(\boldsymbol{W},\boldsymbol{X}) = \Phi^{-1}\left[\xi(\boldsymbol{W},\boldsymbol{X})/\left\{1+\xi(\boldsymbol{W},\boldsymbol{X})\right\}\right] + \text{sign}\left\{R_U(\boldsymbol{W},\boldsymbol{X})R_Y(\boldsymbol{W},\boldsymbol{X})\right\}$. Thus, given $\boldsymbol{W}=\boldsymbol{w}$ and $\boldsymbol{X}=\boldsymbol{x}$, $E'=e'$ is completely determined so that, $\pi_{\text{CVX}}^{\text{opt}}(\boldsymbol{w},\boldsymbol{x}) = \text{sign}\left[\Phi(e')R_Y(\boldsymbol{w},\boldsymbol{x}) + \left\{1-\Phi(e')\right\}R_Z(\boldsymbol{w},\boldsymbol{x})\right]$ which can be seen (after some algebra) to equal $\pi_U^{\text{opt}}(\boldsymbol{w},\boldsymbol{x}) = \text{sign}\left\{R_U(\boldsymbol{w},\boldsymbol{x})\right\}$. ∎

The preceding result states that a latent preference model that assumes a convex utility function is sufficiently expressive provided that the sign of $R_U(\boldsymbol{w}, \boldsymbol{x})$ matches the sign of $R_Y(\boldsymbol{w}, \boldsymbol{x})$ or $R_Z(\boldsymbol{w}, \boldsymbol{x})$; i.e., for any $\boldsymbol{w}, \boldsymbol{x}$ the optimal policy under the unknown utility would recommend a treatment that is satisfactory to patients who care only about outcome Y or care only about outcome Z. Note that violation of this assumption would imply that there exists a pair $(\boldsymbol{x}, \boldsymbol{w})$ such that under U a patient would prefer a treatment that was *worse* on both outcomes which is unrealistic in many settings.

3.2.2 Estimating an Optimal ITR

To construct an estimator of π^{opt} we first express it in terms of the underlying generative model. We assume: (C1) consistency, $(Y, Z) = \{Y^*(A), Z^*(A)\}$; (C2) positivity, there exists $\epsilon > 0$ so that $P(A = a|\boldsymbol{X}, \boldsymbol{W}) \geq \epsilon$ for each a, with probability one; (C3) ignorability, $[\{Y^*(a), Z^*(a)\} : a \in \{-1, 1\}] \perp A\big|\boldsymbol{X}, \boldsymbol{W}$; and (C4) $(A, Y, Z) \perp E\big|\boldsymbol{X}, \boldsymbol{W}$. The first three assumptions are standard (Robins et al. 2000, Zhang et al. 2012b) whereas (C4) can be satisfied provided the conditions of Lemma (3.2.1) hold. Under (C1)-(C3) it can be shown (Schulte et al. 2014) that $\pi^{\text{opt}}(\boldsymbol{x}, \boldsymbol{w}) = \arg\max_{a\in\{-1,1\}} \mathbb{E}\big[\Phi(E)Y + \{1 - \Phi(E)\} Z\big|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w}, A = a\big]$ which, under (C4), can be written as

$$\pi^{\text{opt}}(\boldsymbol{x}, \boldsymbol{w}) = \arg\max_{a\in\{-1,1\}} \Big[\mathbb{E}\left\{\Phi(E)\big|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w}\right\} \mathbb{E}\left(Y\big|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w}, A = a\right) + \left[1 - \mathbb{E}\left\{\Phi(E)\big|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w}\right\}\right] \mathbb{E}\left(Z\big|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w}, A = a\right)\Big].$$

The foregoing expression suggests the following approach for estimating π^{opt}: (i) construct estimators, say $\widehat{Q}_{Y,n}(\boldsymbol{x}, \boldsymbol{w}, a)$ and $\widehat{Q}_{Z,n}(\boldsymbol{x}, \boldsymbol{w}, a)$, of $Q_{Y,n}(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{a}) = \mathbb{E}(Y\big|\boldsymbol{X} =$

$\boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w}, A = a)$ and $Q_Z(\boldsymbol{x}, \boldsymbol{w}, a) = \mathbb{E}(Z|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w}, A = a)$; (ii) postulate a latent preference model linking the unobservable preference E with covariates $\boldsymbol{X}$ and preference questionnaire items $\boldsymbol{W}$ and use this postulated model to construct an estimator, say $\widehat{\mu}_{E,n}(\boldsymbol{x}, \boldsymbol{w})$ of $\mu(\boldsymbol{x}, \boldsymbol{w}) = \mathbb{E}\left\{\Phi(E) \middle| \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w}\right\}$; and (iii) compute $\widehat{\pi}_n(\boldsymbol{x}, \boldsymbol{w}) = \arg\max_{a \in \{-1,1\}} \left[\widehat{\mu}_{E,n}(\boldsymbol{x}, \boldsymbol{w}) \widehat{Q}_{Y,n}(\boldsymbol{x}, \boldsymbol{w}, a) + \{1 - \widehat{\mu}_{E,n}(\boldsymbol{x}, \boldsymbol{w})\} \widehat{Q}_{Z,n}(\boldsymbol{x}, \boldsymbol{w}, a)\right]$. Below we describe an instantiate of this approach that appears to work well in practice and possesses a number of desirable theoretical properties.

To construct estimators of $Q_Y(\boldsymbol{x}, \boldsymbol{w}, a)$ and $Q_Z(\boldsymbol{x}, \boldsymbol{w}, a)$ we postulate linear working models of the form $Q_Y(\boldsymbol{x}, \boldsymbol{w}, a; \psi_Y) = \boldsymbol{x}_{Y,0}^\top \psi_{Y,0} + \boldsymbol{w}_{Y,1}^\top \psi_{Y,1} + a\boldsymbol{x}_{Y,1}^\top \psi_{Y,2} + a\boldsymbol{w}_{Y,1}^\top \psi_{Y,3}$ and $Q_Z(\boldsymbol{x}, \boldsymbol{w}, a; \psi_Z) = \boldsymbol{x}_{Z,0}^\top \psi_{Z,0} + \boldsymbol{w}_{Z,0}^\top \psi_{Z,1} + a\boldsymbol{x}_{Z,1}^\top \psi_{Z,2} + a\boldsymbol{w}_{Z,1}^\top \psi_{Z,3}$, where $\boldsymbol{x}_{\ell,j}$ and $\boldsymbol{w}_{\ell,j}$ for $\ell = Y, Z$ and $j = 0, 1$ are known feature vectors constructed from $\boldsymbol{x}$ and $\boldsymbol{w}$ and ψ_W, ψ_Y are unknown parameter vectors. Note that we assume this linear form for simplicity and in practice this form of the working model should be in line with current literature and previous analyses. Let $\mathbb{P}_n$ denote the empirical measure and define $\widehat{\psi}_{Y,n} = \arg\min_{\psi_Y} \mathbb{P}_n \{Y - Q_Y(\boldsymbol{X}, \boldsymbol{W}, A; \psi_Y)\}^2$ and $\widehat{\psi}_{Z,n} = \arg\min_{\psi_Z} \mathbb{P}_n \{Z - Q_Z(\boldsymbol{X}, \boldsymbol{W}, A; \psi_Z)\}^2$. Subsequently, we construct estimators $Q_Y(\boldsymbol{x}, \boldsymbol{w}, a; \widehat{\psi}_Y)$ and $Q_Z(\boldsymbol{x}, \boldsymbol{w}, a; \widehat{\psi}_Z)$ of $Q_Y(\boldsymbol{x}, \boldsymbol{w}, a)$ and $Q_Z(\boldsymbol{x}, \boldsymbol{w}, a)$.

To develop our latent preference model, we assume (C5) that $E \perp \boldsymbol{X} \big| \boldsymbol{W}$. This assumption simplifies our development and is justified in our application where $\boldsymbol{X}$ does not contain information thought to be informative about patient preferences beyond $\boldsymbol{W}$; however, this assumption is not necessary (De Ayala 2013). We assume that latent patient preferences are connected to items on the questionnaire through a Rasch model (Rasch 1961; 1980) of the form $\text{logit}\left\{P\left(W_j = 1 \middle| E = e\right)\right\} = \beta_{0,j} + \beta_{1,j} e$, $j = 1, \ldots, p$ which is indexed by $\beta = (\beta_{0,1}, \beta_{1,1}, \ldots, \beta_{0,p}, \beta_{1,p})$. Let β^* denote the true parameter value; we use the EM algorithm to construct an estimator, $\widehat{\beta}_n$, of β^* (Rizopoulos 2006).

Given estimator $\widehat{\beta}_n$ and a postulated marginal distribution, say p_e, for the latent preferences, the conditional distribution of E given $\boldsymbol{W} = \boldsymbol{w}$ is proportional to $p(\boldsymbol{w}|e)p_h(e)$ which can be approximated using Metropolis Hastings. Because the conditional distribution of E is only used to construct an estimator of $\mu(\boldsymbol{x}, \boldsymbol{w}) = \mathbb{E}\left\{\Phi(E)\middle|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w}\right\}$ and because we assume that $\mu(\boldsymbol{x}, \boldsymbol{w})$ does not depend on $\boldsymbol{x}$, it is computationally less burdensome to apply a method of moments type estimator. Let $\widehat{e}_n(\boldsymbol{w})$ denote the solution to $\sum_{j=1}^{p} \widehat{\beta}_{n,1,j}\text{expit}\left(\widehat{\beta}_{n,j,0} + \widehat{\beta}_{n,1,j}e\right) = \sum_{j=1}^{p} \widehat{\beta}_{n,1,j}w_j$, where $\text{expit}(u) = \exp(u)/\left\{1 + \exp(u)\right\}$. Subsequently, let $\widehat{\mu}_{E,n}(\boldsymbol{x}, \boldsymbol{w}) = \Phi\left\{\widehat{e}_n(\boldsymbol{w})\right\}$ denote our estimator of $\mu(\boldsymbol{w}, \boldsymbol{x})$. Results provided in Appendix A.1 and A.2 of the Supplemental Material show that this estimator provides qualitatively similar results as Metropolis Hastings while being significantly less computationally intensive.

Theoretical results

Let $e^*(\boldsymbol{w})$ denote the solution to $\sum_{j=1}^{p} \beta^*_{1,j}\text{expit}(\beta^*_{j,0} + \beta^*_{j,1}e) = \sum_{j=1}^{p} \beta^*_{j,1}w_j$. We make the following assumptions

(A1) The number of items satisfies $3 \leq p_n = o(e^n)$.

(A2) The estimator $\widehat{e}_n(\boldsymbol{w})$ converges in probability to $e^*(\boldsymbol{w})$, pointwise for all $\boldsymbol{w}$.

(A3) The estimators $\widehat{Q}_{Y,n}(\boldsymbol{x}, \boldsymbol{w}, a)$ and $\widehat{Q}_{Z,n}(\boldsymbol{x}, \boldsymbol{w}, a)$ converge in probability to $Q_{Y,n}(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{a})$ and $Q_{Z,n}(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{a})$, pointwise for each $\boldsymbol{x}$ and $\boldsymbol{w}$.

The preceding assumptions are rather mild: (A1) can be expected to hold in many applications as it is common to have $p \ll n$; (A2) holds under standard regularlity conditions for methods of moments estimators; and (A3) holds under standard linear modeling assumptions.

The first result establishes consistency of the proposed estimator for the optimal regimen as the sample size diverges but the number of items remains fixed.

Theorem 3.2.2. *Assume (A1)-(A3) and let the number of items, $p_n = p$, be fixed. Then $V_U(\pi^{\text{opt}}) - V_U(\widehat{\pi}_n)$ converges to zero in probability as $n \to \infty$.*

The preceding result is relevant when the number of items is small relative to the number of enrolled subjects, e.g., in the illustrative example presented in Section 3.4 there are $p = 10$ items and $n = 957$ subjects. However, the next result shows that if the number of items is allowed to diverge with the sample size then the estimated optimal regime performs as well as an oracle with access to each patient's individual preference.

Theorem 3.2.3. *Assume (A1)-(A3) and suppose $p_n \to \infty$ as $n \to \infty$.*
Define $\pi^{\text{oracle}}(x, e) = \arg\max_a \mathbb{E}(U|X = x, E = e, A = a)$ denote the optimal policy if *patient preferences were known. Then $V_U(\widehat{\pi}_n) - V_U(\pi^{\text{oracle}})$ converges to zero in probability as $n \to \infty$.*

The proof of the preceding result involves, as an intermediate step, proving consistency of the Rasch model for individual patient preferences when the number of items is nearly exponential in the number of patients. To our knowledge, this is new result in item response theory that may be of independent interest.

3.3 Simulation Study

In this section we examine the finite sample performance of the proposed estimation method in terms of the average value obtained. Additional simulations comparing the method of moments estimator with Metropolis Hastings are provided in Appendix A.2 of the Supplemental Material.

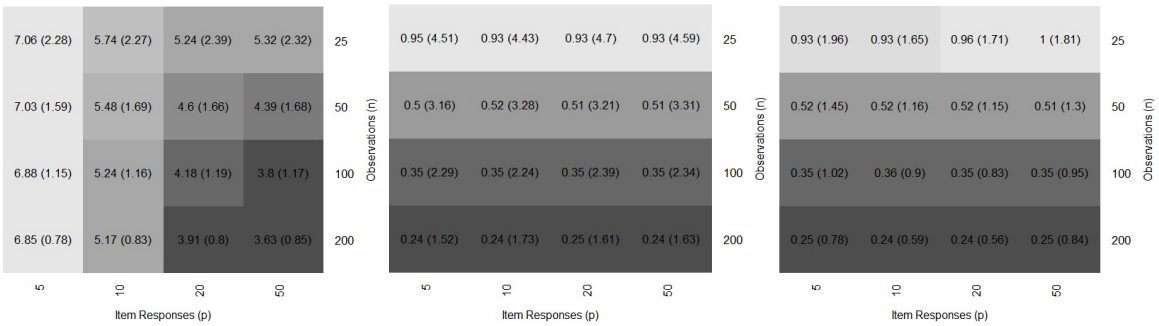
3.3.1 Data Generating Model

We consider the following class of generative models. Latent preferences are drawn as $i.i.d.$ so that $E \sim \text{Normal}(0,1)$. Responses to items are drawn $i.i.d.$ so that $W_j \sim \text{Bernoulli}\{\text{expit}(\beta_{0,j} + \beta_{1,j}E)\}$; the true beta parameters are chosen to ensure that when $\widehat{E}_n(\boldsymbol{W})$ was regressed on $\boldsymbol{W}$ using a probit regression model, the Nagelkerke's R^2 was equal to 0.5. The treatment, covariates, and outcome data were drawn $i.i.d.$ so that: $A \sim Unif\{-1,1\}$, $\boldsymbol{X} \sim N_5(0, I_5)$, $Y = \boldsymbol{X}^\intercal\psi_{00} + A\boldsymbol{X}^\intercal\psi_{01} + \epsilon$ and $Z = \boldsymbol{X}^\intercal\psi_{10} + A\boldsymbol{X}^\intercal\psi_{11} + \delta$ where $\epsilon, \delta \sim N(0,1)$, $\psi_{00} = (2.5, .2, .25, -.7, -2.5, 2.4)$, $\psi_{01} = (1.7, -2.3, 4.5, 6, -7.3, -1.6)$, $\psi_{10} = t + q^*\psi_{00}$ and $\psi_{11} = t + q^*\psi_{01}$. We set $q = -2$ and $t = 3$ so that the outcomes favor different treatments about 85% of the time; other settings provide qualitatively similar results and are provided in the Supplemental Material.

3.3.2 Simulation Results

The simulations were each repeated $s = 500$ times for all combinations of $n = 25, 50, 100, 200$ and $p = 5, 10, 20, 50$. Let $Q_U(\boldsymbol{x}, \boldsymbol{w}, a) = \mathbb{E}\left(U \middle| \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w}, A = a\right)$ and $\widehat{Q}_U(\boldsymbol{x}, \boldsymbol{w}, a) = \widehat{\mu}_{E,n}(\boldsymbol{x}, \boldsymbol{w})\widehat{Q}_{Y,n}(\boldsymbol{x}, \boldsymbol{w}, a) + \{1 - \widehat{\mu}_{E,n}(\boldsymbol{x}, \boldsymbol{w})\}\,\widehat{Q}_{Z,n}(\boldsymbol{x}, \boldsymbol{w}, a)$ to be the proposed estimator of $Q_U(\boldsymbol{x}, \boldsymbol{w}, a)$. In our first simulation experiment we compute the average squared difference between the estimated value function and the true value function of the optimal regime. The results are displayed in Figure 3.1. When we compare the average squared difference between $\widehat{V}_U(\widehat{\pi}_n) = \mathbb{P}_n \max_a \widehat{Q}_U(\boldsymbol{X}, \boldsymbol{W}, a)$ and $V_U(\pi^{\text{opt}})$, it can be seen that the quality of the estimated value improves as either n or p increases. As expected, we can see from the averaged squared distance between $\widehat{V}_Y(\widehat{\pi}_n) = \mathbb{P}_n \max_a \widehat{Q}_Y(\boldsymbol{X}, \boldsymbol{W}, a)$ and $V_Y(\pi^{\text{opt}})$ that the quality of the approximation improves as n increases but is insensitive to changes in p. We see an analogous pattern in the average difference between $\widehat{V}_Z(\widehat{\pi}_n)$ and $V_Z(\pi^{\text{opt}})$.

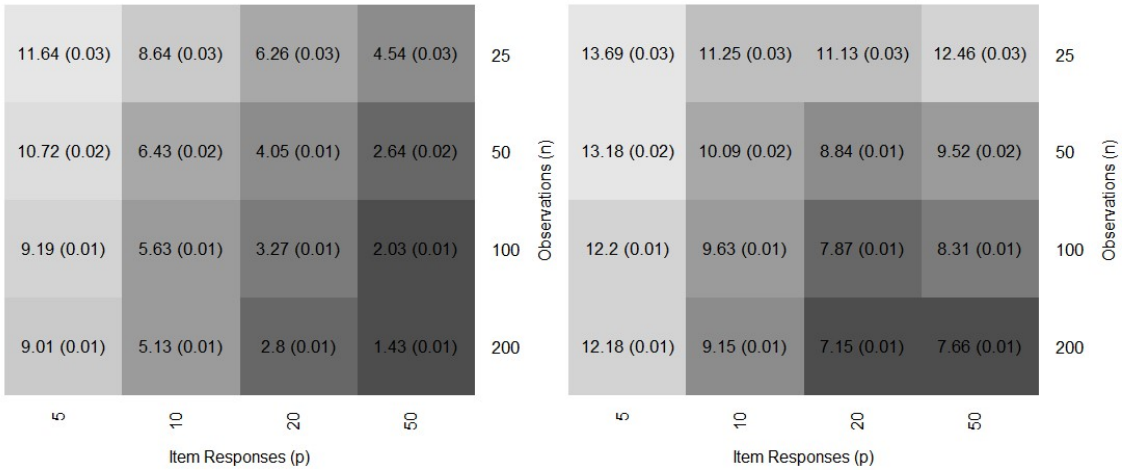
Figure 3.1: Averaged squared difference: $\widehat{V}_U(\widehat{\pi}_n)$, $\widehat{V}_Y(\widehat{\pi}_n)$ and $\widehat{V}_Z(\widehat{\pi}_n)$



a : $\widehat{V}_U(\widehat{\pi}_n)$ and $V_U(\pi^{\text{opt}})$ b : $\widehat{V}_Y(\widehat{\pi}_n)$ and $V_Y(\pi^{\text{opt}})$ c : $\widehat{V}_Z(\widehat{\pi}_n)$ and $V_Z(\pi^{\text{opt}})$

We also compared $\widehat{\pi}_n$ with π^{opt} and π^{oracle} in terms of treatment recommendations. Figure 3.2 shows the average percentage of disagreement of $\widehat{\pi}_n$ with π^{opt} and π^{oracle}. It can be see that the level of agreement is generally high and, as anticipated by the theoretical results in the preceding section, the agreement improves as n and p increase.

Figure 3.2: Average percent disagreement: $\widehat{\pi}_n$, π^{opt} and π^{oracle}



a : Average disagreement between $\widehat{\pi}_n$ and π^{opt} b : Average disagreement between $\widehat{\pi}_n$ and π^{oracle}

3.4 Case Study

The Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) schizophrenia trial was designed to compare new antipsychotic drugs to conventional ones in a randomized, controlled, double blinded, multi-phase trial (Lieberman et al. 2005, Stroup et al. 2003, Davis et al. 2003; 2011). The trial was intended for patients who were already being treated for schizophrenia but who might benefit from a medicinal change. The patients not only received antipsychotic treatments, but were also offered psychosocial treatment with their families. In the first phase of the trial patients were randomized to one of 5 medications: 4 of which were atypical antipsychotics (olanzapine, quetiapine, risperidone and ziprasidone) and one conventional antipsychotic (perphenazine). The primary study analysis (Lieberman et al. 2005) compared all treatments individually, and identified substantial variability amongst the atypicals. However, for the purposes of this illustration, we dichotomize treatment into two groups: atypical antipsychotics and perphenazine. Patients who discontinued their treatment would move to phase 2 and would be randomized to another treatment according to the protocol. The structure of the first phase of this trial makes it an ideal application for the method presented in this paper. At baseline, patients answered a 10 question, binary-response assessment, which we use as a measure of the patient's preference across two outcomes: (i) efficacy measured using the Positive and Negative Syndromes Scale (PANSS) (Kay et al. 1987); and (ii) side effect burden measured as the sum of side effect and adverse event indicators. Detailed descriptions of these outcomes are provided in Appendix A.3 of the Supplemental Material. Outcomes were measured at the end of the first phase, or throughout the phase for adverse events. The end of the first phase occurred when the patient underwent a treatment change (at the discretion of the patient and clinician) or 18 months, whichever came first.

The patient preference information was collected using a 10 question Drug Attitude Inventory (Hogan et al. 1983) assessment that each patient answered at the baseline. We opted to exclude one question ("I take medications of my own free choice") because it cannot be interpreted as a trade-off between two outcomes of interest. The remaining 9 questions are as follows and were coded so that (1) favors efficacy and (0) favors side effect burden.

1. "For me, the good things about medication outweigh the bad" Yes (1) No (0)
2. "I feel weird, like a 'zombie', on medication" Yes (0) No (1)
3. "Medications make me feel more relaxed" Yes (1) No (0)
4. "Medications make me feel tired and sluggish" Yes (0) No (1)
5. "I take medications only when I'm sick" Yes (0) No (1)
6. "I feel more normal on medication" Yes (1) No (0)
7. "It is unnatural for my mind and body to be controlled by medications" Yes (0) No (1)
8. "My thoughts are clearer on medication" Yes (1) No (0)
9. "By staying on medications I can prevent being sick" Yes (1) No (0)

We selected tailoring covariates based on clinical expertise and prior analyses (Shortreed and Moodie 2012, Shortreed et al. 2014). The variables included in our analysis are gender, age, race (white, black, other), BMI, diastolic blood pressure, systolic blood pressure, baseline clinical severity of schizophrenia, age at first antipsychotic medication, and any antipsychotic medication they were taking at the baseline (olanzapine, quetiapine, risperidone, ziprasidone, perphenazine, haldol, or any long-term injectable antipsychotic).

The estimated optimal treatment allocations for each treatment by outcome is

displayed in Table 3.1. While efficacy appears to favor atypical antipsychotics, side effect burden tends to favor perphenazine. As expected, the composite outcome occupies a middle ground between the two marginal outcomes; the optimal treatment for the composite outcome and efficacy are the same 92% of the time, the optimal treatment for the composite outcome and side effect burden are the same 64% of the time, and the optimal treatment for efficacy and side effect burden are the same 56% of the time. Another view of these data is presented in Table 3.2 which shows the fraction of overlap between the estimated optimal regime using the proposed estimator and the optimal regime based only on efficacy or side effect burden.

Table 3.1: Treatment recommendations

	Perphenazine	Atypical Antipsychotics
$\widehat{\pi}_Y$: Efficacy	36%	64%
$\widehat{\pi}_Z$: Side Effect Burden	67%	33%
$\widehat{\pi}_n$: $\Phi(E)Y + \{1 - \Phi(E)\} Z$	41%	59%

Table 3.2: Percent of agreement in treatment recommendations

	$\widehat{\pi}_Y$: Efficacy	$\widehat{\pi}_Z$: Side Effect Burden
Both recommend perphenazine	35%	36%
Only $\widehat{\pi}_n$ recommends perphenazine	6%	5%
Only $\widehat{\pi}_n$ recommends atypical antipsychotic	2%	31%
Both recommend atypical antipsychotic	57%	28%

3.5 Discussion

Clinical and intervention scientists must often balance multiple, possibly competing outcomes when designing a personalized treatment plan for a patient. We combined a latent trait model with Q-learning to incorporate individual patient preferences into an evidence-based treatment plan formalized as a treatment regime. The proposed estimator is consistent for the optimal regime and possesses an oracle property if the number of items on the preference questionnaire is allowed to diverge with sample size; incidentally, to establish these results we developed new theory for Rasch model that may be of interest beyond the application we consider here. We view the proposed methodology as an important first step toward mathematical formalizations of shared decision making in the context of precision medicine.

There are number of interesting and important extensions of the proposed methodology. The extension to settings with multiple treatment periods is of particular interest. This extension is challenging as a patient's preferences may change over time in response to treatment received and interim outcomes experienced, e.g., a patient who experiences severe side effects may develop a strong aversion to them in the future. Furthermore, in applications where interventions are delivered on a fine time scale, e.g., mobile-health, one must judiciously choose the timetable for administering the preference questionnaire to balance information gain with patient burden.

CHAPTER 4: ESTIMATING DYNAMIC TREATMENT STRATEGIES BY INTEGRATING PATIENT PREFERENCES

4.1 Introduction

The clinical treatment of patients with chronic disease is evolving every day. Patient care now focuses on determining which treatment(s) should be administered, who they should be administered to, and at which stage in the treatment plan they should be administered (Collins and Varmus 2015). Individual characteristics such as prognostic information, treatment history and familial data are taken into consideration when estimating a time dependent set of treatment rules called dynamic treatment regimes (DTRs). DTRs serve as a set of treatment decisions that evolve through time determined by patient progress and prognostic variables with the goal of optimizing a patient's long term outcome (Collins et al. 2014). These rules take advantage of the fact that treatments have different levels of efficacy in tandem with other treatments. This ensures they identify which course of treatment is the best overall as opposed to myopically determining which treatment is the best at each stage. The statistical goal is to mathematically mimic how clinicians treat patients in practice (Collins et al. 2007). The type of data that is needed to estimate DTRs is most efficiently collected from a sequential multiple assignment randomization trial. These trials specialize in using a small sample size to efficiently collect information on multiple treatment paths (Kidwell 2014). Therefore, it is often important to develop an accurate estimate of the optimal treatment strategy that can accommodate a reduced sample size.

These individual characteristics can also include patient preference information regarding how to weigh the importance of two competing clinical outcomes (Edwards and Elwyn 2009). There are multiple ways to collect preference information. Patients can directly choose parameters that summarize their preferences. Unfortunately, this often requires formal training, which is not feasible in most treatment settings (Brennan 1998, Braziunas 2006, Lizotte et al. 2012a). Alternatively, since this information is considered unobservable, the patient can answer a series of questions on a questionnaire and, from this, the patient's preference can be estimated (Embretson and Reise 2013). These questions can require a patient to choose from binary responses or rank a set of ordered responses that correspond to one of the competing outcomes. While there is a multitude of research on how to collect and even summarize latent patient information, there has been little work on how to incorporate it when estimating DTRs.

The method presented here estimates optimal DTRs for competing outcomes, such as efficiency versus toxicity, in line with the patient's preferences concerning these two outcomes. The idea is that one treatment may well favor one outcome and the other treatment favor the other. It is important to determine how, when, and to whom these should be allocated. The ideal situation is to accurately elicit patient preferences and integrate them into a composite outcome that creates a linear trade off between the two clinical outcomes. We chose to use an item response theory approach to estimate a patient's preferences by analyzing the set of dichotomous responses to carefully constructed questions using a latent trait model (Rizopoulos 2006). The optimal treatment rule is determined as that which provides the maximum expected utility given the patient's history, assuming that the best treatment is assigned at all subsequent stages. The preference information is updated at each treatment stage based on how their preferences have changed and their overall contentment with their health status. Section 3 presented the basis for this work in the single stage scenario.

Our method extends that to the multiple stage scenario using a reinforcement learning technique called Q-learning (Watkins and Dayan 1992). For the purposes of our work, the method is only presented for the two-stage setting since it is mathematically trivial to extend this to 3 or more decision stages. In a simulation study, the estimated optimal dynamic treatment rule is compared with the true optimal dynamic treatment rule and the optimal static treatment rule for varying sample sizes and number of questions in the questionnaire. The static treatment rule assumes patient preferences do not change over time and uses the preference information only collected at the baseline; otherwise, the rest of the estimation procedure is the same. While this method has the desirable asymptotic properties, it is also shown to be quite accurate with smaller sample sizes and a reasonably sized preference questionnaire.

4.2 Estimating the Optimal DTR

4.2.1 Framework

Assume the observed data are

$$\{(\boldsymbol{X}_i^1, \boldsymbol{W}_i^1, A_i^1, B_i^1, Y_i^1, Z_i^1, \ldots, \boldsymbol{X}_i^T, \boldsymbol{W}_i^T, A_i^T, B_i^T, Y_i^T, Z_i^T, \boldsymbol{W}_i^{T+1})\}_{i=1}^n$$

which comprise n iid trajectories of the form $(\boldsymbol{X}^1, \boldsymbol{W}^1, A^1, B^1, Y^1, Z^1, \ldots,$ $\boldsymbol{X}^T, \boldsymbol{W}^T, A^T, B^T, Y^T, Z^T, \boldsymbol{W}^{T+1})$. Then, $\boldsymbol{X}^t \in \mathbb{R}^m$ indicates the patient covariate information collected preceding assigning the treatment at time t; $\boldsymbol{W}^t \in \{0,1\}^p$ denotes responses to the itemized questionnaire administered before assign the treatment at time t; $A^t \in \mathbb{A}^t$ denotes the treatment assigned at time t; $Y^t \in \mathbb{R}$ is the first outcome measured after receiving treatment at time t; $Z^t \in \mathbb{R}$ is the second outcome measured after receiving treatment at time t. $B^t \in \{0,1\}$ denotes an indicator that

the patient is content with the outcome observed after receiving treatment at time t. We include a measure of the patient's contentment because it allows us to calibrate how well the assigned treatment is aligning with the patient's preferences and their observed outcome. Let $\boldsymbol{H}^t$ denote the patient's history, or the information available to the decision maker before treatment assignment at time t so that $\boldsymbol{H}^1 = \boldsymbol{X}^1$ and $\boldsymbol{H}^t = \{\boldsymbol{H}^{t-1}, A^{t-1}, B^{t-1}, Y^{t-1}, Z^{t-1}, \boldsymbol{X}^t\}$ for $t = 2, \ldots, T$.

In the context treatment of schizophrenia, our motivating example, $\boldsymbol{X}$ denotes patient prognostic information such as demographic variables; $\boldsymbol{W}$ measures subject's responses on a set of questions concerning the two competing outcomes (severity of psychotic symptoms and aggregate side effect burden) ; A denotes two treatments, one of which is associated with higher alleviation of depression severity but higher side effects and the other is associated with lower alleviation of depression severity but also lower side effects; B indicates if the patient is satisfied with their overall wellbeing; Y denotes the severity of the depression; and Z denotes a measure of aggregate side-effect severity.

We define the set of dynamic treatment rules as $\boldsymbol{\pi} = (\pi^1, \ldots, \pi^T)$. At stage t, the function $\pi^t : \operatorname{dom} \boldsymbol{W}^t \times \operatorname{dom} \boldsymbol{H}^t \to \operatorname{dom} A^t$ is a map from the patient's itemized response and history to the recommended treatment rule. This means that for π^t a patient with questionnaire responses $\boldsymbol{W}^t = \boldsymbol{w}^t$ and history $\boldsymbol{H}^t = \boldsymbol{h}^t$ would be recommended treatment $\pi^t(\boldsymbol{w}^t, \boldsymbol{h}^t)$. We assume that for each time, t, a patient's preferences are summarized by $E^t \in \mathbb{R}$. E^t denotes the patients value of outcomes Y^t compared to Z^t and are assumed to change across time. The utility function that represents a trade off between the outcomes in line with the patient's preferences is $u(y^t, z^t, e^t)$. Then, a patient with preference $E^t = e^t$ would prefer the outcome set (y^t, z^t) over $(\tilde{y}^t, \tilde{z}^t)$ if $u(y^t, z^t, e^t) > u(\tilde{y}^t, \tilde{z}^t, \tilde{e}^t)$. Let $V(\pi^t) = \mathbb{E}^{\pi}[u(Y^t, Z^t, E^t)]$ denote the value function which is the expected utility under treatment π^t. Then, the treatment rule $\pi^{t,opt}$ is

said to be optimal if $V(\pi^{t,opt}) \geq V(\pi^t) \quad \forall \quad \pi^t$. The goal is to estimate $\pi^{t,opt}$ for all t. Let us denote $Y^{t,*}(a^t)$ and $Z^{t,*}(a^t)$ as the potential outcomes for Y^t, Z^t at stage t under treatment $a^t \in \{-1, 1\}$ (Rubin 1978). Then denote the utility function as $u(Y^{t,*}(a^t), Z^{t,*}(a^t), E^{t,*})$ and the value function as

$$V_u(\pi^t) = \mathbb{E}\left[\sum_{a^t \in \{-1,1\}} u\left\{Y^{t,*}(a^t), Z^{t,*}(a^t), E^t\right\} 1\{\pi\left(\boldsymbol{W}^t, \boldsymbol{H}^t\right) = a^t\}\right].$$

The optimal treatment at time t, $\pi_u^{t,\text{opt}}$, is that which satisfies $V_u\left(\pi_u^{t,\text{opt}}\right) \geq V_u\left(\pi_u^t\right)$ for all π_u^t.

We propose the following additive parametric model for the utility function, $u(y, z, e) = \Sigma_{t=1}^T u(y^t, z^t, e^t; \rho^t)$, which is indexed by an unknown parameter $\rho^t \in \mathbb{R}^d$. The optimal treatment rule can be characterized in terms of the underlying generative model. For any ρ^t, define

$$Q^T(\boldsymbol{h}^T, a^T) \triangleq \mathbb{E}\{\Sigma_{t=1}^T u(Y^t, Z^t, E^t; \rho^t) | \boldsymbol{H}^T = \boldsymbol{h}^T, A^T = a^T\}$$

as the Q-function for the last follow-up stage T. Then $\pi^{T,opt}(\boldsymbol{h}^T) = \text{argmax}_{a^T} Q^T(\boldsymbol{h}^T, a^T)$. The backwards recursive Q-functions for $T = T-1, T-2, \ldots, 1$ are defined as

$$Q^t(\boldsymbol{h}^t, a^t) \triangleq \mathbb{E}\{\max_{a^{t+1}} Q^{t+1}(\boldsymbol{h}^{t+1}, a^{t+1}) | \boldsymbol{H}^t = \boldsymbol{h}^t, A^t = a^t\}$$

where $\pi^{t,opt}(\boldsymbol{h}^t) = \text{argmax}_{a^t} Q^t(\boldsymbol{h}^t, a^t)$. Provided $\widehat{Q}^t$ is an estimator of Q^t, then the optimal treatment rule for $t = 1, \ldots, T$ is $\widehat{\pi}_n^t(h^t) = \text{argmax}_{a^t} \widehat{Q}_n^t(\boldsymbol{h}^t, a^t)$.

4.2.2 Methodology

We assume a linear trade off between the outcomes because it is reasonable to assume a patient would prefer one outcome to the other by a certain magnitude. Hence, we set the utility function to be $u(Y^t, Z^t, E^t; \rho^t) = \Phi(E^t)Y^t + (1 - \Phi(E^t))Z^t\rho^t$ so that the weight is a function of the patient preferences. Note that $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and is used as a normalizing factor to ensure $\Phi(\boldsymbol{E}^t) \in [0, 1]$. In this function, ρ^t represents either a penalty or reward that is commensurate with contentment. ρ^t can take any value greater than zero. By increasing or decreasing the magnitude of the second outcome's contribution to the utility function, we can better control for the patient's feelings about the results of their previous treatment. To construct our estimator, we make the following assumptions: (C1) consistency, $(Y^t, Z^t) = \{Y^{t,*}(A^t), Z^{t,*}(A^t)\}$, which implies the individuals observed outcome is the potential outcomes associated with the observed treatment; (C2) positivity, $P(A = a|\boldsymbol{X}, \boldsymbol{W}) \geq \epsilon$ for some $\epsilon > 0$ and each a , with probability one, which implies that there is a positive probability of receiving all possible treatments; (C3) ignorability,
$[\{Y^{t,*}(a^t), Z^{t,*}(a^t)\} : a^t \in \{-1, 1\}] \perp A^t \big| \boldsymbol{H}^t, \boldsymbol{W}^t$; and (C4) $(A^t, Y^t, Z^t) \perp E^t \big| \boldsymbol{H}^t, \boldsymbol{W}^t$, which reasonably implies that the patient's preferences are independent of the the assigned treatment and the outcomes. The first three assumptions are standard (Robins et al. 2000, Zhang et al. 2012b). (C4) can be satisfied as shown in section 3.2.2. Under

these assumptions the optimal policy can be written as

$$\pi^{\mathrm{t,opt}}(\boldsymbol{h}^t, \boldsymbol{w}^t) =$$
$$\arg\max_{a^t in\{-1,1\}} \Big[\mathbb{E}\left\{\Phi(E^t)\middle|\boldsymbol{H}^t = \boldsymbol{h}^t, \boldsymbol{W}^t = \boldsymbol{w}^t\right\} \mathbb{E}\left(Y^t\middle|\boldsymbol{H}^t = \boldsymbol{h}^t, \boldsymbol{W}^t = \boldsymbol{w}^t, A^t = a^t\right)$$
$$+ \left[1 - \mathbb{E}\left\{\Phi(E^t)\middle|\boldsymbol{H}^t = \boldsymbol{h}^t, \boldsymbol{W}^t = \boldsymbol{w}^t\right\}\right] \mathbb{E}\left(Z^t\middle|\boldsymbol{H}^t = \boldsymbol{h}^t, \boldsymbol{W}^t = \boldsymbol{w}^t, A^t = a^t\right) \widehat{\rho}^t \Big].$$

The following describes the approach we took to estimating the sequence of dynamic treatment rules $\boldsymbol{\pi}^{\mathrm{opt}} = (\pi^{1,\mathrm{opt}}, \ldots, \pi^{\mathrm{T,opt}})$:

1. Develop the posterior distribution model of the latent preference information, E^t, given the patient's history, $\boldsymbol{H}^t$, contentment of outcomes after treatment, B^t, and responses to the preference questionnaire, $\boldsymbol{W}^t$, for all t. Use this posterior model to construct and estimate the patient's estimated conditional latent preference at the final stage, $t = T$, as $\widehat{\eta}_{E^T,n}(\boldsymbol{h}^T, \boldsymbol{w}^T)$ where $\eta(\boldsymbol{h}^T, \boldsymbol{w}^T) = \mathbb{E}\left\{\Phi(\boldsymbol{E}^T)\middle|\boldsymbol{H}^T = \boldsymbol{h}^T, \boldsymbol{W}^T = \boldsymbol{w}^T\right\}$.

2. Construct estimators of the conditional distribution of each of the outcomes given the patient's history, H^T, responses to the questionnaire, $\boldsymbol{W}^T$, and treatment, A^T, where $Q_{Y^T,n}(\boldsymbol{h}^T, \boldsymbol{w}^T, a^T) = \mathbb{E}(Y^T|\boldsymbol{H}^T = \boldsymbol{h}^T, \boldsymbol{W}^T = \boldsymbol{w}^T, A^T = a^T)$ and $Q_{Z^T}(\boldsymbol{h}^T, \boldsymbol{w}^T, a^T) = \mathbb{E}(Z^T|\boldsymbol{H}^T = \boldsymbol{h}^T, \boldsymbol{W}^T = \boldsymbol{w}^T, A^T = a^T)$. Note that we assume a linear form of the conditional distribution for simplicity, but this can easily be generalized to any other form in practice.

3. Define

$$\widehat{Q}_n^T(\boldsymbol{h}^T, \boldsymbol{w}^T, a^T) = \widehat{\eta}_{E^T,n}(\boldsymbol{h}^T, \boldsymbol{w}^T)\widehat{Q}_{Y^T,n}(\boldsymbol{h}^T, \boldsymbol{w}^T, a^T)$$
$$+ \{1 - \widehat{\eta}_{E^T,n}(\boldsymbol{h}^T, \boldsymbol{w}^T)\}\widehat{Q}_{Z^T,n}(\boldsymbol{h}^T, \boldsymbol{w}^T, a^T)\rho^T$$

and $\widehat{\pi}^{T,opt} = \arg\max_{a^T \in \{-1,1\}} \widehat{Q}_n^T(\boldsymbol{h}^T, \boldsymbol{w}^T, a^T)$.

4. Using Q-learning, repeat steps (i)-(iii) for $t = T-1, T-2, \ldots, 1$ with the exception that the Q-function is defined as follows:

$$\widehat{Q}_n^{t,*}(\boldsymbol{h}^t, \boldsymbol{w}^t, a^t) = \widehat{\eta}_{E^t,n}(\boldsymbol{h}^t, \boldsymbol{w}^t)\widehat{Q}_{Y^t,n}(\boldsymbol{h}^t, \boldsymbol{w}^t, a^t) + \{1 - \widehat{\eta}_{E^t,n}(\boldsymbol{h}^t, \boldsymbol{w}^t)\}\widehat{Q}_{Z^t,n}(\boldsymbol{h}^t, \boldsymbol{w}^t, a^t)\rho^t$$

$$\widehat{Q}_n^t(\boldsymbol{h}^t, \boldsymbol{w}^t, a^t) = \widehat{Q}_n^{t,*}(\boldsymbol{h}^t, \boldsymbol{w}^t, a^t) + \widehat{Q}_n^{t+1}(\boldsymbol{h}^{t+1}, \boldsymbol{w}^{t+1}, a^{t+1} = \widehat{\pi}^{t+1,opt})$$

We assume the preference and utility are related to the observed data through three models, the item response model, the contentment model and the preference evolution model as follows:

- (Item Response Model) $\text{logit}P(W_j^t = 1 | \boldsymbol{H}^t = \boldsymbol{h}^t, E^t = e^t) = \gamma_{j,0}^t + \gamma_{j,1}^t e^t$
- (Contentment Model) $\text{logit}P(B^t = 1 | \boldsymbol{H}^t = \boldsymbol{h}^t, A^t = a^t, Y^t = y^t, Z^t = z^t, E^{t+1} = e^{t+1}) = \zeta_0 + \zeta_1 u(y^t, z^t, e^{t+1}; \rho^t)$
- (Preference Evolution Model) $E^1 \sim N(0,1)$; $E^{t+1} = \mu(y^t, z^t; \tau) + \tau_1 + e^t + \tau_2 a^t e^t + \epsilon^t, \quad \epsilon^t \sim N(0,1)$

The form of μ is assumed to be known as $\mu(Y^t, Z^t; \tau) = \tau_{00} + \tau_{01} y^t + \tau_{02} z^t$. Notice that the contentment depends on the future preference information, e^{t+1}. This is because at the time B^t is collected the patient is describing their contentment with the outcomes (Y^t, Z^t) in terms of their current preferences and not the one that preceded the assignment of A^t. The estimates of β are found from the latent trait model; estimates of τ^t are obtained using the following linear regression model: $\mathbb{E}(E^t) = \tau_0^t + \tau_1^t y^{t-1} + \tau_2^t z^{t-1} + \tau_3^t e^{t-1} + \tau_4^t a^{t-1} e^{t-1}$; estimates of ζ^t, ρ^t are obtained using logistic regression of the following model: $\text{logit}P(B^t = 1 | \boldsymbol{H}^t = \boldsymbol{h}^t, A^t = a^t, Y^t = y^t, Z^t =$

$z^t, E^{t+1} = e^{t+1}) = \zeta_1^t + \zeta_2^t[\Phi(e^{t+1})y^t + \{1 - \Phi(e^{t+1})\}z^t\rho^t]$. Given these estimated parameters, the posterior distribution of the latent preference is defined as

$$P(E^t|H^t) \propto P(B^{t-1}|E^t, Y^{t-1}, Z^{t-1})P(W^t|E^t)P(E^t|Y^{t-1}, Z^{t-1}).$$

From this, a Metropolis Hastings algorithm can be used to estimate the posterior mean where the estimator of the trade-off weight is $\widehat{\eta}_{E^T,n}(\boldsymbol{h}^T, \boldsymbol{w}^T) = \Phi(\widehat{e}^t)$.

We define linear working models of the form $Q_{Y^T}(\boldsymbol{h}^T, \boldsymbol{w}^T, a^T; \gamma_Y) = \boldsymbol{h}^T\gamma_{0,0} + a^T\boldsymbol{h}^T\gamma_{0,1}$ and $Q_{Z^T}(\boldsymbol{h}^T, \boldsymbol{w}^T, a^T; \gamma_Z) = \boldsymbol{h}^T\gamma_{1,0} + a^T\boldsymbol{h}^T\gamma_{1,1}$ and obtain parameter estimates of $\widehat{\gamma}_{0,0}, \widehat{\gamma}_{0,1}, \widehat{\gamma}_{1,0}, \widehat{\gamma}_{1,1}$ using simple linear regression. Then, $\widehat{Q}_{Y^T,n}(\boldsymbol{h}^T, \boldsymbol{w}^T, a^T)$ and $\widehat{Q}_{Z^T,n}(\boldsymbol{h}^T, \boldsymbol{w}^T, a^T)$ be the fitted estimators such that

$$\widehat{Q}_n^T(\boldsymbol{h}^T, \boldsymbol{w}^T, a^T) = \Phi(\widehat{e}^T)\widehat{Q}_{Y^T,n}(\boldsymbol{h}^T, \boldsymbol{w}^T, a^T) + [1 - \Phi(\widehat{e}^T)]\widehat{Q}_{Z^T,n}(\boldsymbol{h}^T, \boldsymbol{w}^T, a^T)\widehat{\rho}^T.$$

For $t < T$ the structure of the linear working models for $\widehat{Q}_{Y^t,n}, \widehat{Q}_{Z^t,n}$ is the same. The Q-function for the composite outcome is similar such that $\widehat{Q}_n^t(\boldsymbol{h}^t, \boldsymbol{w}^t, a^t) = \widehat{Q}_n^{t,*}(\boldsymbol{h}^t, \boldsymbol{w}^t, a^t) + \widehat{Q}_n^{t+1}(\boldsymbol{h}^{t+1}, \boldsymbol{w}^{t+1}, a^{t+1} = \widehat{\pi}^{t+1,opt})$.

4.3 Simulation Study

4.3.1 Data Generating Model

For the simulation study, we generated the data as to mimic what would be collected in a clinical trial. Each treatment, A_i^t, is randomly generated as 1 or -1. The covariates, $\boldsymbol{X}$, are assumed to be observed at all T stages and are correlated between stages for each subject. λ and κ are randomly chosen to be used for all simulations.

The covariance of $\boldsymbol{X}_i$ is defined using the following ARMA(1,1) structure:

$$\Sigma_{X_i} = \begin{bmatrix} 1 & \lambda & \lambda^*\kappa & \lambda^*\kappa^3 & \dots & \lambda^*\kappa^{t-1} \\ \lambda & 1 & \lambda & \lambda^*\kappa^2 & \dots & \lambda^*\kappa^{t-2} \\ \vdots & & & \ddots & & \vdots \\ \lambda^*\kappa^{t-1} & \lambda^*\kappa^{t-2} & \lambda^*\kappa^{t-3} & \lambda & \dots & 1 \end{bmatrix}$$

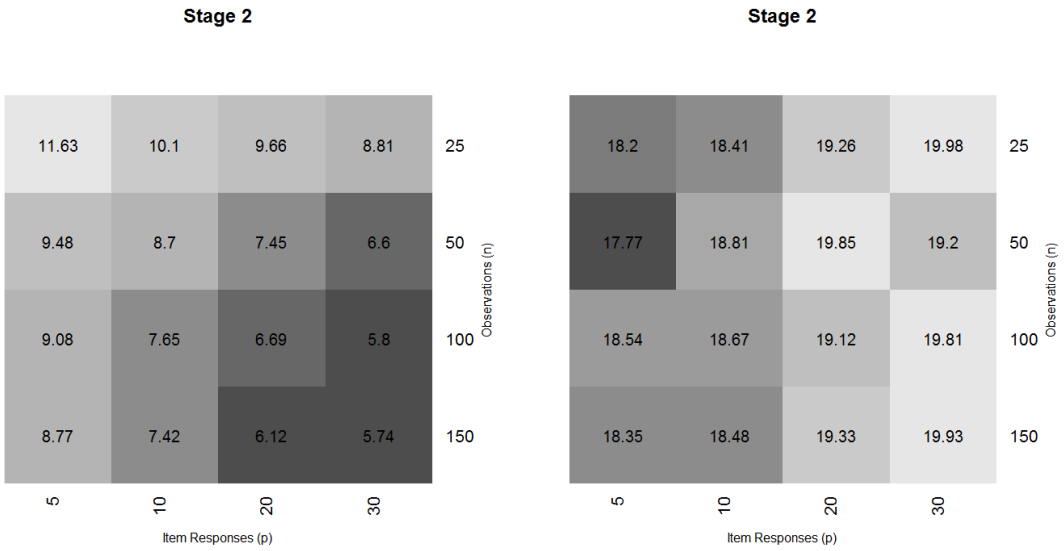
For each of the m=5 covariates, $\boldsymbol{x}_{i,t} = \{x_{i,1}, x_{i,2}, \dots, x_{i,T}\} \sim MVN(\underline{0}_T, \Sigma_{X_i})$ At stage $t > 1$, each of the covariates experienced a positive or negative "drift" depending on if they received treatment $A = 1$ or $A = -1$ at stage 1. The purpose of this drift is to mimic the effect of the treatment on each of the covariates at the subsequent stage. At the first stage, we let $\epsilon_i^1 \sim N(0,1)$ and $\delta_i^1 \sim N(0,1)$ be random noise, where $\boldsymbol{\epsilon}^1 = \{\epsilon_i^1, \dots, \epsilon_n^1\}$ and $\boldsymbol{\delta}^1 = \{\delta_i^1, \dots, \delta_n^1\}$. Let Y^1 and Z^1 be the outcomes at the first stage. Then, $Y^1 = \boldsymbol{X}^1\gamma_{0,0} + A^1\boldsymbol{X}^1\gamma_{0,1} + \boldsymbol{\epsilon}^1$ and $Z^1 = \boldsymbol{X}^1\gamma_{1,0} + A^1\boldsymbol{X}^1\gamma_{1,1} + \boldsymbol{\delta}^1$. At the second stage, we let $\epsilon_i^2 \sim N(0,1)$ and $\delta_i^2 \sim N(0,1)$ be random noise, where $\boldsymbol{\epsilon}^2 = \{\epsilon_i^2, \dots, \epsilon_n^2\}$ and $\boldsymbol{\delta}^2 = \{\delta_i^2, \dots, \delta_n^2\}$. Let Y^2 and Z^2 be the outcomes at the second stage and $G^1 = (\boldsymbol{X}^1, A^1, Y^1, Z^1)$. Then, $Y^2 = \boldsymbol{X}^2\gamma_{0,0} + A^2\boldsymbol{X}^2\gamma_{0,1} + G^1\gamma_{0,2} + \boldsymbol{\epsilon}^2$ and $Z^1 = \boldsymbol{X}^1\gamma_{1,0} + A^2\boldsymbol{X}^2\gamma_{1,1} + G^1\gamma_{1,2} + \boldsymbol{\delta}^2$. For the latent preference information, E, we assume $E^1 \sim N(0,1)$ at the first stage and $E^t = \tau_1^t Y^{t-1} + \tau_2^t Z^{t-1} + \tau_3^t E^{t-1} + \tau_4^t A^{t-1} E^{t-1} + \epsilon$ for all subsequent stages. The responses the the itemized questionnaire, $\boldsymbol{W}$, are assumed to follow a binomial distribution such that $\boldsymbol{W}_{i,j}^t | E_i^t \sim \text{Bernoulli}(\text{expit}\{\beta_{0,j,k} + \beta_{1,j,k} E_i^t\})$. To generate the measure of contentment after each stage, B, let $\rho^t > 0, \zeta_1^t, \zeta_2^t \in \mathbb{R}$ and $u^t = \Phi(E^t)Y^{t-1} + (1 - \Phi(E^t))Z^{t-1}\rho^t$. We also assume B^t follow a Bernoulli distribution such that $B^t \sim \text{Bernoulli}(\text{expit}\{\zeta_1^t + \zeta_2^t u^t\})$.

4.3.2 Simulation Results

For all combinations of $n = 25, 50, 100, 200$ and $p = 5, 10, 20, 50$, the simulations were repeated $s = 500$ times. In all instances, we assume there are only $t = 2$ treatment stages, but the steps of Q learning can easily be extended to more than 2 stages.

Figure 4.1 shows the percent of simulations the true and estimated dynamic treatment rule differ at the second stage and the percent of simulations the estimated dynamic and static treatment rules differ for at the second stage. When looking at the true versus estimated dynamic treatment rule heatmap (on the left), we see that for stage 2 the method becomes more accurate as n and p increase. While there is still increased accuracy, the magnitude of the increase decreases from 100 observations to 150 observations, which may suggest it is sufficient to only include 100 observations. We also see a small increase in the magnitude of the estimator's accuracy between 20 items and 30 items, which may also suggest that it is sufficient to only include 20 items in the preference questionnaire. Turning to the estimated optimal dynamic versus static treatment rule (on the right), there appears to be more disagreement as p increases which implies that as we increase the number of items in the questionnaire the static estimator and the dynamic estimator are favoring different outcomes. As expected, there is not much evidence of a trend across n. This is because the difference between the dynamic and static estimator is the estimated latent preference information, $\Phi(\widehat{e})$, which is dependent on p, not n.

Figure 4.1: Average percent disagreement: $\widehat{\pi}_D^2$, π_T^2 and $\widehat{\pi}_S^2$

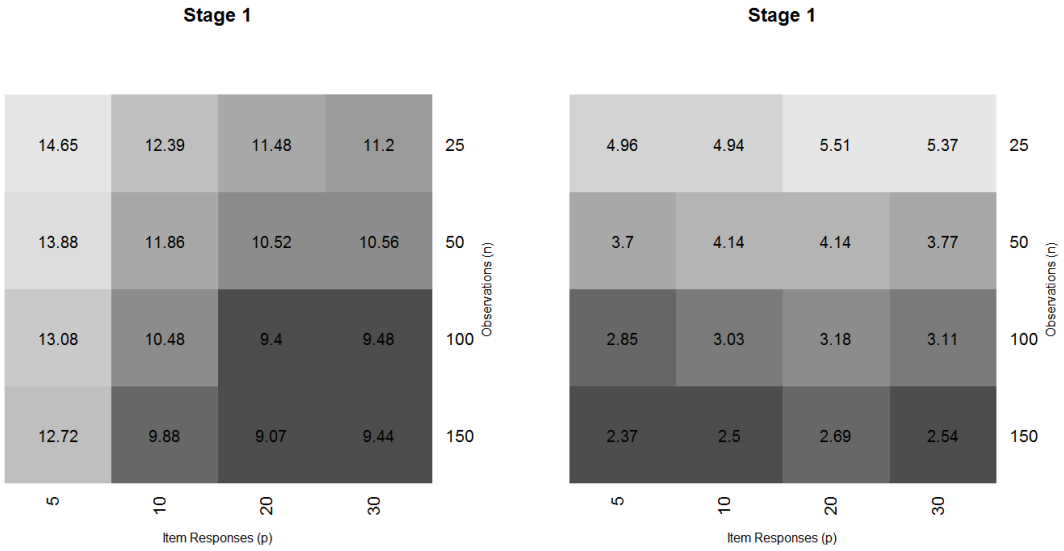


a : Average disagreement between $\widehat{\pi}_D^2$ and π_T^2 b : Average disagreement between $\widehat{\pi}_D^2$ and $\widehat{\pi}_S^2$

Figure 4.2 shows the percent of simulations the true and estimated dynamic treatment rule differ at the first stage and the percent of simulations the estimated optimal dynamic and static treatment rules differ for at the first stage. In looking at the true versus estimated optimal treatment rule heatmap (on the left), we see that at stage 1, the method becomes more accurate as n and p increase. We note that when n = 150 and p = 30 there is a slight increase in disagreement, which may imply that that many observations combined with items does not provide much additional information. We also notice that there is more disagreement in stage 1 than stage 2, which may be attributed to the mechanisms of Q-learning. This is because in all the steps where $t < T$, the Q function incorporates the expected utility of using the optimal treatment at the future stage, which introduces some variability. Turning to the estimated optimal dynamic versus static treatment rule (on the right), there is substantially less disagreement at stage 1 than at stage 2. This is likely because the estimated preference from stage 1 is used for stage 2 in the static estimator. It appears there is slightly more disagreement

between the dynamic and static rules for smaller sample sizes, but not much difference across items.

Figure 4.2: Average percent disagreement: $\widehat{\pi}_D^1$, π_T^1 and $\widehat{\pi}_S^1$



a : Average disagreement between $\widehat{\pi}_D^1$ and π_T^1 b : Average disagreement between $\widehat{\pi}_D^1$ and $\widehat{\pi}_S^1$

4.4 Discussion

We have presented a new method to estimate optimal dynamic treatment plans that consider patient preferences about a set of 2 competing outcomes. The pre-specified utility function serves as the composite outcome and is a linear combination of the preference information and the two competing outcomes. This new outcome is used in Q-learning to estimate which treatment is optimal at each stage, assuming that the best treatment is given at subsequent stages. In each stage, the Q-function estimates the utility of each treatment given the patient's personal and treatment history and assigns the treatment that elicits the highest utility. The latent preference information is estimated using an itemized questionnaire and whether the patient was content with their status after receiving their previous treatment. This is done using item response

theory, specifically the Rasch probability model, and utilizes a Metropolis Hastings algorithm to estimate the conditional latent preference information. Under the given assumptions, the method proves accurate, even for smaller sample sizes and a small number of items on the questionnaire. This method provides a way for clinicians to tailor the treatment regime to each patient based on their preferences as well as how they are progressing at each point in time. It also takes the whole sequential treatment plan into account as opposed to estimating the treatment at each stage on its own.

There are a handful of limitations to this work that can expanded upon in the future to provide meaningful progress in the field of precision medicine. The composite outcome is a linear function of the predicted outcomes and the latent preference estimate. This could be improved upon by using a nonparametric or semi-parametric composite outcome. We also only looked at two competing outcomes. It would strengthen this line of work to be able to account for 3 or more competing outcomes. Balancing 3 outcomes requires a more developed form of the latent trait model for polytomous data (that could be linear, or nonlinear, in form). It would also be extremely beneficial to test this method on real data, which unfortunately does not exist at the moment. With more developed methodology to choose from, it is our hope that data will become available to test this method in the future. This method is a new tool researchers can use to encourage the implementation of new, patient centered, sequential multiple treatment assignment trials in precision medicine.

CHAPTER 5: NONPARAMETRIC INCORPORATION OF PATIENT PREFERENCES FOR INDIVIDUALIZED TREATMENT RULE ESTIMATION

5.1 Introduction

To effectively and efficiently treat patients, particular care must be taken when determining which treatments should be administered and to whom. The population for any specific disease is vast and diverse and therefore it is not sufficient to apply a one size fits all approach to health care. Precision medicine is the practice of specifying healthcare plans, such as treatment allocation, to patients in a way that is best for them and patients like them. It is a broad term used to describe targeted therapies based a patient's genetic information, prognostic variables and past history. The goal is to be able to determine what patients should be getting what treatments and if they should be receiving the treatments at all (Jameson and Longo 2015, Sox et al. 2008). In clinical settings where the patient or doctor is provided with a selection of treatment options, it may be desirable to include a patient's preference into decision making. Consider the treatment of a mental health disorder as a tangible example. We can imagine a scenario where managing the illness involves understanding how the patient feels about two or more outcomes, such as efficacy and side effect burden. In this example, as well as in other therapeutic areas, it is difficult, if not impossible, to quantify how the patient feels about the two competing outcomes at any instance in time. One solution is to survey a set of responses from the patient (or clinician) that can be used to mathematically infer their personal preference evaluation using item response theory (Embretson and Reise

2013). This estimation of the patient preferences can then be included when estimating the optimal treatment rule.

When seeking to estimate a function that maps current patient information to the treatment space, we formalize it as an individualized treatment rule (ITR). An optimal ITR is a mathematically dependent treatment plan that mimics how patient's are actually treated by their physician. ITRs seek to improve a clinician's well educated guess by assigning the treatment that provides the best expected outcome in the population for patient's experiencing similar characteristics (Lavori and Dawson 2014). While there has been a lot of work done for estimating ITRs, there has been limited progress in developing ITRs that weigh competing outcomes. One way to measure the trade off between the two competing outcomes is through a linear trade-off. The question becomes how do we obtain a weight to define this trade off? Furthermore, how can we define the relationship between the item response data and the patient preference if we want a model more flexible than the Rasch model (Rasch 1961)?

In this paper, we propose a nonparametric approach to solving this problem. It is desirable to use nonparametric estimation because it requires more general assumptions, which are more realistic in a practical application. During a clinical evaluation, a patient is instructed to fill out an itemized questionnaire where the responses correspond to one of two competing outcomes. This serves as parsimonious collection of questions with binary response options that are used to evaluate how the patient feels about these outcomes. These outcomes are often associated with the treatment options such that one treatment elicits a positive effect in one outcome, a negative effect in the other outcome, and vice versa. Once this item response information is collected, we use it to estimate the patient's latent preference. For our purposes, the relationship between the item responses and the preference information is assumed to be nonparametric. To find an approximation, we use a piecewise linear splines that are constrained to be

monotonic (Villalobos and Wahba 1987) and from this create a posterior estimate of the patient's preferences given the item responses. This spline model is often referred to as a broken stick function. A spline is a data smoothing technique that connects a set of linear functions at connectors called knots. These knots ensure continuity while the function in between the knots is fit for just that range of values. Splines are a good nonparametric estimation technique because they allow for flexibility as well as simplicity of implementation. The parameters within the spline function can then be estimated subject to a set of necessary constraints using nonlinear programming. These preferences are normalized and then used to define the patient's ideal linear trade off between the two outcomes. To do this, we employ monotonic splines once again to obtain the relationship between the patient's satisfaction with the outcomes after receiving the treatment. This more accurately captures the patient's feelings on how to weigh the two outcomes. From this, the linear utility function serves as quantity we wish to optimize. The optimal treatment rule is that which maximizes the patient's utility function. The method is tested in a simulation study which compares the accuracy of the estimated optimal treatment rule for a sequence of varying samples sizes and number of response items in the questionnaire.

5.2 Optimal Nonparametric ITRs with Patient Preferences

5.2.1 Framework and Notation

Assume the observed data is defined as $\{(\boldsymbol{W}_i, \boldsymbol{X}_i, A_i, Y_i, Z_i, B_i)\}_{i=1}^n$ which are n independent and identically distributed observations from $(\boldsymbol{W}, \boldsymbol{X}, A, Y, Z, B)$. Here, $\boldsymbol{X} \in \mathbb{R}^m$ represents patient information preceding the treatment assignment; $\boldsymbol{W} \in \{0,1\}^p$ represents responses to itemized questionnaire; $A \in \mathbb{A}$ represents the treatment assignment; $Y \in \mathbb{R}$ is the first outcome measured after receiving the assigned treatment;

$Z \in \mathbb{R}$ is the second outcome measured after receiving the assigned treatment; $B \in \{0, 1\}$ represents an indicator that that the patient is content with the outcome observed after receiving treatment. B is only collected during the clinical trial that serves as the training data set to determine the set of treatment rules. This is to validate how satisfied the patient is with the results of the treatment assignment. In this method, the responses to the itemized questionnaire, $\boldsymbol{W}$ are used to estimate the latent preference, $E \in \mathbb{R}$.

Under the ITR, π, a patient with itemized responses $\boldsymbol{W} = \boldsymbol{w}$ and covariates $\boldsymbol{X} = \boldsymbol{x}$ would be recommended treatment $\pi(\boldsymbol{w}, \boldsymbol{x})$. This is because π is a map from the itemized responses and patient covariates to a recommended treatment as. $\pi : \operatorname{dom} \boldsymbol{W} \times \operatorname{dom} \boldsymbol{X} \to \operatorname{dom} A$. We define the utility function, $U(Y, Z; E)$, as a composition of the observed outcomes and the latent preference information, and is intended to serve as the ultimate outcome of interest in our estimation. We note that the latent preference information, E, ensures ordering such that, if $U(y, z; e) > U(y', z'; e)$ patient's with $E = e$ prefer outcomes to $\{Y = y, Z = z\}$ to $\{Y = y', Z = z'\}$. Letting $Y^*(a)$ and $Z*(a)$ represent potential outcomes for assigned treatment a, the potential utility function under a is then $U(Y^*(a), Z^*(a); E)$ (Rubin 1978). For any π, the potential utility can then be defined as $V_U(\pi) = \mathbb{E}\left[\sum_{a \in \{-1,1\}} U\left\{Y^*(a), Z^*(a); E\right\} 1_{\pi\left(\boldsymbol{W}, \boldsymbol{X}\right)=a}\right]$. Hence, the optimal rule is $\pi_U^{\text{opt}} \in \arg\max_\pi V_U(\pi)$ for all ITRs π (Zhao et al. 2012).

To define the nonparametric utility function, we employ monotonic splines (He and Shi 1998). Let $G = \Phi(E)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable so that $G \in [0, 1]$. Then, define the utility of the form $U(Y, Z; E) = h(G)Y + \{1 - h(G)\}Z$, where h is a monotone spline that maps from $[0, 1]$ to $[0, 1]$. In this form, the spline will encompass all reasonable utilities. Each of the splines are constrained to be equal at the interior points, have slopes are positive, and such that the first linear function has an intercept of 0 and the last linear function

is constrained such that the summation of the slope and intercept is 1.

If we define:

$$\begin{aligned}
R_Z(\boldsymbol{w},\boldsymbol{x}) &= \mathbb{E}\left\{Z^*(1)\middle|\boldsymbol{W}=\boldsymbol{w},\boldsymbol{X}=\boldsymbol{x}\right\}-\mathbb{E}\left\{Z^*(-1)\middle|\boldsymbol{W}=\boldsymbol{w},\boldsymbol{X}=\boldsymbol{x}\right\},\\
R_Y(\boldsymbol{w},\boldsymbol{x}) &= \mathbb{E}\left\{Y^*(1)\middle|\boldsymbol{W}=\boldsymbol{w},\boldsymbol{X}=\boldsymbol{x}\right\}-\mathbb{E}\left\{Y^*(-1)\middle|\boldsymbol{W}=\boldsymbol{w},\boldsymbol{X}=\boldsymbol{x}\right\},\\
R_U(\boldsymbol{w},\boldsymbol{x}) &= \mathbb{E}\left[U\left\{Y^*(1),Z^*(1);E\right\}\middle|\boldsymbol{W}=\boldsymbol{w},\boldsymbol{X}=\boldsymbol{x}\right]\\
&\quad -\mathbb{E}\left[U\left\{Y^*(-1),Z^*(-1);E\right\}\middle|\boldsymbol{W}=\boldsymbol{w},\boldsymbol{X}=\boldsymbol{x}\right],
\end{aligned}$$

then, it can be shown that $\pi_U^{\text{opt}}(\boldsymbol{w},\boldsymbol{x}) = \text{sign}\{R_U(\boldsymbol{w},\boldsymbol{x})\}$ (Qian and Murphy 2011). Since $h(\Phi(E))$ is a linear combination of E, from Lemma 3.1 in section 3.2.1, we can assume the utility function is sufficiently defined as long as the sign of $R_U(\boldsymbol{w},\boldsymbol{x})$ is the same as the sign of $R_Y(\boldsymbol{w},\boldsymbol{x})$ or $R_Z(\boldsymbol{w},\boldsymbol{x})$. This means that the patient would be content with the optimal policy if if they only care about one of the outcomes.

5.2.2 Estimation

To construct our estimator, π^{opt}, we make three standard causal inference assumptions (Robins et al. 2000, Zhang et al. 2012b): (C1) consistency, $(Y,Z) = \{Y^*(A),Z^*(A)\}$; (C2) positivity, for each a there exists $\epsilon > 0$ such that $P(A=a|\boldsymbol{X},\boldsymbol{W}) \geq \epsilon$; (C3) ignorability, $[\{Y^*(a),Z^*(a)\} : a \in \{-1,1\}] \perp A\big|\boldsymbol{X},\boldsymbol{W}$. The consistency assumption implies that the individual's observed outcome is the potential outcome associated with the observed exposure; the positivity assumption implies there is a positive probability of receiving all possible treatments; the ignorability assumption implies that patients receiving all possible treatments have equal distributions of experiencing the potential outcomes. We further assume (C4) $(A,Y,Z) \perp E\big|\boldsymbol{X},\boldsymbol{W}$,

which states that the treatment assignment and responses are independent of a patient's preference. This important assumption is reasonable in the single stage scenario because the treatments are assigned randomly and if outcomes were affected by the preferences, treatments would be obsolete. Under these stated assumptions, it can be shown (Schulte et al. 2014) that the optimal individualized treatment can then be defined in terms of the predefined utility function:

$$\begin{aligned}
\pi^{\text{opt}}(\boldsymbol{x},\boldsymbol{w}) &= \arg\max_{a\in\{-1,1\}} \mathbb{E}\big[U(Y,Z;E)\big|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{W}=\boldsymbol{w},A=a\big] \\
&= \arg\max_{a\in\{-1,1\}} \mathbb{E}\big[h(G)Y+\{1-h(G)\}\,Z\big|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{W}=\boldsymbol{w},A=a\big] \\
&= \arg\max_{a\in\{-1,1\}} \Big[\mathbb{E}\left\{h(G)\big|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{W}=\boldsymbol{w}\right\}\mathbb{E}\left(Y\big|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{W}=\boldsymbol{w},A=a\right) \\
&\quad +\left[1-\mathbb{E}\left\{h(G)\big|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{W}=\boldsymbol{w}\right\}\right]\mathbb{E}\left(Z\big|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{W}=\boldsymbol{w},A=a\right)\Big] \\
&= \arg\max_{a\in\{-1,1\}} \Big[\mu(\boldsymbol{x},\boldsymbol{w})Q_Y(\boldsymbol{x},\boldsymbol{w},\boldsymbol{a})+[1-\mu(\boldsymbol{x},\boldsymbol{w})]\,Q_Z(\boldsymbol{x},\boldsymbol{w},a)\Big],
\end{aligned}$$

where $Q_Y(\boldsymbol{x},\boldsymbol{w},a) = \mathbb{E}(Y\big|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{W}=\boldsymbol{w},A=a)$, $Q_Z(\boldsymbol{x},\boldsymbol{w},a) = \mathbb{E}(Z|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{W}=\boldsymbol{w},A=a)$ and $\mu(\boldsymbol{x},\boldsymbol{w}) = \mathbb{E}\left\{h(G)\,\big|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{W}=\boldsymbol{w}\right\}$. Now let $\widehat{Q}_{Y,n}(\boldsymbol{x},\boldsymbol{w},a)$, $\widehat{Q}_{Z,n}(\boldsymbol{x},\boldsymbol{w},a)$ and $\widehat{\mu}_n(\boldsymbol{x},\boldsymbol{w})$ be respective estimators of $Q_Y(\boldsymbol{x},\boldsymbol{w},\boldsymbol{a})$, $Q_Z(\boldsymbol{x},\boldsymbol{w},a)$ and $\mu(\boldsymbol{x},\boldsymbol{w})$. Then,

$$\widehat{\pi}_n(\boldsymbol{x},\boldsymbol{w}) = \arg\max_{a\in\{-1,1\}} \left[\widehat{\mu}_n(\boldsymbol{x},\boldsymbol{w})\widehat{Q}_{Y,n}(\boldsymbol{x},\boldsymbol{w},a)+\{1-\widehat{\mu}_n(\boldsymbol{x},\boldsymbol{w})\}\,\widehat{Q}_{Z,n}(\boldsymbol{x},\boldsymbol{w},a)\right]$$

This implies that estimation $\pi^{\text{opt}}(\boldsymbol{x},\boldsymbol{w})$ can be broken down into three estimators: $\widehat{Q}_{Y,n}(\boldsymbol{x},\boldsymbol{w},a)$, $\widehat{Q}_{Z,n}(\boldsymbol{x},\boldsymbol{w},a)$ and $\widehat{\mu}_n(\boldsymbol{x},\boldsymbol{w})$

To define $\widehat{Q}_{Y,n}(\boldsymbol{x},\boldsymbol{w},a)$ and $\widehat{Q}_{Z,n}(\boldsymbol{x},\boldsymbol{w},a)$, we further assume $Y,Z \perp \boldsymbol{W}\big|\boldsymbol{X}$. This

means that the itemized response information contained in $\boldsymbol{W}$ as it relates to the outcomes is sufficiently captured in the covariates $\boldsymbol{X}$ and need not be included in our estimator. Then, we construct linear working models as $Q_Y(\boldsymbol{x}, \boldsymbol{w}, a; \gamma_Y) = \boldsymbol{x}_{Y,0}^\top \gamma_{Y,0} + a\boldsymbol{x}_{Y,1}^\top \gamma_{Y,1}$ and $Q_Z(\boldsymbol{x}, \boldsymbol{w}, a; \gamma_Z) = \boldsymbol{x}_{Z,0}^\top \gamma_{Z,0} + a\boldsymbol{x}_{Z,1}^\top \gamma_{Z,1}$, where $\boldsymbol{x}_{\ell,j}$ for $\ell = Y, Z$ and $j = 0, 1$ are known feature vectors constructed from $\boldsymbol{x}$ and γ_Y, γ_Z are unknown parameter vectors. Note that we assume a linear working model, but this form is note required and was just assumed for simplicity. Let $\widehat{\gamma}_{Y,n}$ and $\widehat{\gamma}Z, n$ be the maximum likelihood estimates such that $\widehat{\gamma}_{Y,n} = \arg\min_{\gamma_Y} \mathbb{E}\left\{Y - Q_Y(\boldsymbol{X}, \boldsymbol{W}, A; \gamma_Y)\right\}^2$ and $\widehat{\gamma}_{Z,n} = \arg\min_{\gamma_Z} \mathbb{E}\left\{Z - Q_Z(\boldsymbol{X}, \boldsymbol{W}, A; \gamma_Z)\right\}^2$. Then, utilizing these estimates, we find that $\widehat{Q}_{Y,n}(\boldsymbol{x}, \boldsymbol{w}, a; \widehat{\gamma}_Y)$ and $\widehat{Q}_{Z,n}(\boldsymbol{x}, \boldsymbol{w}, a; \widehat{\gamma}_Z)$ are the estimators of $Q_Y(\boldsymbol{x}, \boldsymbol{w}, a)$ and $Q_Z(\boldsymbol{x}, \boldsymbol{w}, a)$.

Constructing and estimator $\widehat{\mu}_n(\boldsymbol{x}, \boldsymbol{w})$ is sufficiently more difficult because E is latent and therefore unobservable. The first step in defining this estimator is finding an estimate for $\Phi(E)$. To develop a latent preference model, we assume that $E \perp \boldsymbol{X} \big| \boldsymbol{W}$. As we will see, this assumption simplifies the construction of the estimator and is reasonable if $\boldsymbol{X}$ does not contain any additional information about patient preferences beyond that contained in $\boldsymbol{W}$. Since W_j is a binary response variable, we can assume that, conditional on the unobserved e, W_j are independent Bernoulli random variables. We define the generating model for $\boldsymbol{W}$ as $P(W_j | E = e) = \frac{exp\{f_j(e)\}}{1+exp\{f_j(e)\}}$, where f_j is constrained to be a monotone increasing function for $j = 1, \ldots, p$. To solve this we employ an EM algorithm. In the estimation step we create starting values for E using the standard Rasch model from traditional item response theory (Rasch 1961; 1980), which we discussed at length in Section 3.2.2. We note that in the EM algorithm these starting values will be replaced with estimates of $\widehat{e_n}$ as the algorithm progresses iteratively.

With these starting values, e^0, we start the maximization step by finding $\widehat{f_j}(e^0)$ using monotone spline smoothing. Let K_E be the number of knots such that the knots,

$\{t_{1,e}, \ldots, t_{k_E,e}\}$ are quantiles of e^0. For computational and asymptotic simplicity, we impose piecewise linear splines and let $\pi_j(e^0) = \{\pi_{j,1}(e^0), \pi_{j,2}(e^0), \ldots, \pi_{j,K_E}(e^0)\}^T$ be the set of splines of order 2. We then estimate f_j by $\widehat{f}_{j,n}(e^0) = \pi_j(e^0)^T$. Define the linear component as $\pi_{j,k} = a_{j,k} + b^*_{j,k}e^0$ for $k = 1, \ldots, K_E$. The linear parameters $a_{j,k}, b_{j,k}$ for a fixed j and all $k = 1, \ldots, K_E$ are simultaneously estimated by solving the following equation using nonlinear programming (Ghalanos and Theussl 2015, Ye 1987):

$$\pi_{j,k}(e^0) = a_{j,k} + b^*_{j,k}e^0 \quad \forall \quad k \in \{1, \ldots, K_G\}$$
$$\text{where} \quad \text{logit}\left\{P\left(W_j = 1 \middle| E = e\right)\right\} = \pi_j(e(^0))$$
$$\text{s.t.}$$
$$a_{j,k} > 0, \quad b_{j,k} > 0 \quad \forall \quad k \in \{1, \ldots, K_E\}$$
$$a_{j,k} + b_{j,k}t_{k,E} = a_{j,k+1} + b_{j,k+1}t_{k,E} \quad \forall \quad k \in \{1, \ldots, K_E\}.$$

Because the W_j are independent for $j \in 1, \ldots, p$, this process is repeated for all j.

Let $\beta_E = \{a_{1,1}, b_{1,1}, \ldots, a_{1,K_E}, b_{1,K_E}, \ldots, a_{p,1}, b_{p,1}, \ldots, a_{p,K_E}, b_{p,K_E}\}$. Once we have estimates $\widehat{\beta}_E$ for β_E, we can create an estimator of e, $\widehat{e}_n$. Given $\widehat{\beta}_E$ and a marginal distribution for the latent patient preferences, the conditional distribution of E given $\boldsymbol{W} = \boldsymbol{w}$ is proportional to $p(\boldsymbol{w}|h)p_h(h)$. This can be approximated using a Metropolis Hastings algorithm. However, because we assume that $\mu(\boldsymbol{x}, \boldsymbol{w})$ does not depend on $\boldsymbol{x}$, it is less computationally intensive to apply a method of moments estimator. We let $\widehat{e}_n$ denote the solution to $\sum_{j=1}^p \widehat{b}_{j,l}\text{expit}\left(\widehat{a}_{j,l} + \widehat{b}_{j,l}e\right) = \sum_{j=1}^p \widehat{b}_{j,l}w_j \ \forall \ l = 1, \ldots, K_E$, where $\text{expit}(u) = \exp(u)/\{1 + \exp(u)\}$. This estimator for e, $\widehat{e}_n$, provides similar estimates as the Metropolis Hastings algorithm while being significantly less computationally

burdensome.

Finally, recall that we only collect the contentment information, B, to develop our rule. We need to incorporate that information into the estimate of $h(\widehat{G})$, which serves as linear trade off weight. To do this, we assume the distribution of B is dependent on the composite outcome, which is a function combination of the outcomes and the preference information as $U(Y, Z; \widehat{e}_n) = h(\widehat{G})Y + \{1 - h(\widehat{G})\}Z$. Hence, we are interested in is an estimator for $h(\widehat{G})$ that optimizes that patients contentment. We assume that $\text{logit}\{P(B = 1)\} = r\{U(Y, Z; \widehat{e}_n)\}$, where r is a monotone increasing function. Let $\widehat{G} = \Phi(\widehat{e}_n)$, then, similar to what we have previously done, we fit this model using monotone spline smoothing except this time we simultaneously fit $h(\widehat{G})$ and $r\{U(Y, Z; \widehat{e}_n)\}$ as piecewise linear splines. Let K_G be the number of knots such that the knots $\{t_{1,G}, \ldots, t_{K_G,G}\}$ are quantiles of $\widehat{G}$. Under these specifications, we impose splines $\pi_h(\widehat{G})$ and $\pi_r\{U(Y, Z; \widehat{e}_n)\}$ where $\pi_h(\widehat{G}) = \{\pi_{h,1}(\widehat{G}), \pi_{h,2}(\widehat{G}), \ldots, \pi_{h,K_G}(\widehat{G})\}$ and $\pi_r\{U(Y, Z; \widehat{e}_n)\} = [\pi_{r,1}\{U(Y, Z; \widehat{e}_n)\}, \pi_{r,2}\{U(Y, Z; \widehat{e}_n)\}, \ldots, \pi_{r,K_G}\{U(Y, Z; \widehat{e}_n)\}]$ for $h(\widehat{G})$ and $r\{U(Y, Z; \widehat{e}_n)\}$ respectively. Then, we solve the nonlinear function:

$$\pi_{r,k}\{U(Y,Z;\widehat{e}_n)\} = \delta_{1,k} + \delta_{2,k}U(Y,Z;\widehat{e}_n) \quad \forall \quad k \in \{1,\ldots,K_G\}$$

$$\text{where} \quad \text{logit}\{P(B=1)\} = \pi_r\{U(Y,Z;\widehat{e}_n)\}$$

$$\text{and} \quad U(Y,Z;\widehat{e}_n) = h(\widehat{G})Y + \{1-h(\widehat{G})\}Z,$$

$$\pi_{h,k}(\widehat{G}) = \alpha_{1,k} + \alpha_{2,k}\widehat{G} \quad \forall \quad k \in \{1,\ldots,K_G\}$$

$$\text{where} \quad h(\widehat{G}) = \pi_h(\widehat{G})$$

$$\text{s.t.}$$

$$\delta_{2,k} > 0, \quad \alpha_{2,k} > 0 \quad \forall \quad k \in \{1,\ldots,K_G\}$$

$$\alpha_{1,1} = 0$$

$$\alpha_{1,K_G} + \alpha_{2,K_G} = 1$$

$$\alpha_{1,k} + \alpha_{2,k}t_{k,G} = \alpha_{1,k+1} + \alpha_{2,k+1}t_{k,G} \quad \forall \quad k \in \{1,\ldots,K_G\}$$

$$\delta_{1,k} + \delta_{2,k}t_{k,G} = \delta_{1,k+1} + \delta_{2,k+1}t_{k,G} \quad \forall \quad k \in \{1,\ldots,K_G\}$$

Then, the estimator of $\mu(\boldsymbol{x},\boldsymbol{w})$ is $\widehat{\mu}_n(\boldsymbol{x},\boldsymbol{w}) = \widehat{h}(\widehat{G})$.

5.3 Simulation Study

The results of the simulation study are included for 3 knots, which has been shown to be sufficient by He and Shi (1998). For completeness, similar results for 5 knots are included in the Appendix B.1. We chose 3 knots because additional knots do not provide enough additional information to justify a more complex model with a burdensome computation.

5.3.1 Setup and Assumptions

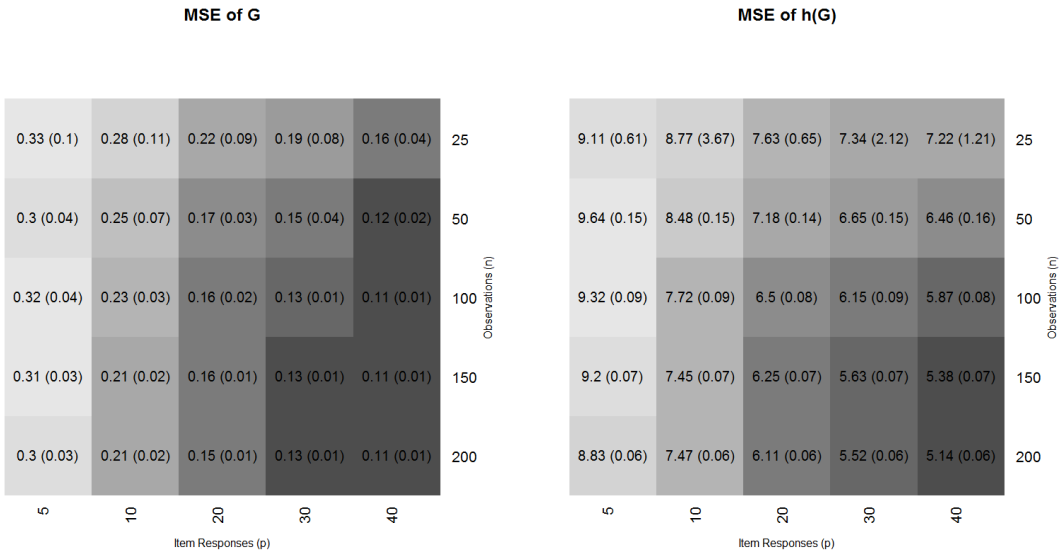
The simulation study considered the following class of models. The patient preferences are assumed to be *i.i.d.* from a standard normal distribution such that $E \sim \mathrm{N}(0,1)$. We designate $h(G)$ to be a piecewise linear spline that is data dependent. With $U(Y,Z;E) = h(G)Y + \{1-h(G)\}Z$, the contentment observation, B, is assumed to be *i.i.d.* from a Bernoulli distribution such that $B \sim \text{Bernoulli}\big\{\text{expit}\big(r\{U(Y,Z;e)\}\big)\big\}$, where $r\{U(Y,Z;e)$ is a piecewise linear spline. The itemized responses are assumed to be *i.i.d.* from a Bernoulli distribution such that $W_j \sim \text{Bernoulli}\big\{\text{expit}\big(f(E)\big)\big\}$, where $f(E)$ is a piecewise linear spline. The treatment, covariates, and outcome data were drawn *i.i.d.* so that: $A \sim Unif\{-1,1\}$, $\boldsymbol{X} \sim N_5(\mu, I_5)$, $Y = \boldsymbol{X}^\top\gamma_{Y,0} + A\boldsymbol{X}^\top\gamma_{Y,1} + \epsilon$ and $Z = \boldsymbol{X}^\top\gamma_{Z,0} + A\boldsymbol{X}^\top\gamma_{Z,1} + \delta$ where $\epsilon, \delta \sim N(0,1)$, $\gamma_{00} = (2.5, .2, .25, -.7, -2.5, 2.4)$, $\gamma_{01} = (1.7, -2.3, 4.5, 6, -7.3, -1.6)$, $\gamma_{10} = t + q^*\gamma_{00}$ and $\gamma_{11} = t + q^*\gamma_{01}$. We set $q = -2$ and $t = 3$ so that the outcomes favor different treatments about 80% of the time.

5.3.2 Simulation Results

The simulations were each repeated $s = 500$ times for all combinations of $n = 25, 50, 100, 150,\ 200$ and $p = 5, 10, 20, 30, 40$. To check the accuracy of this method, it is beneficial to dissect the estimator and check the accuracy of each component. Since $\widehat{Q}_{Z,n}(\boldsymbol{x}, \boldsymbol{w}, a)$ and $\widehat{Q}_{Z,n}(\boldsymbol{x}, \boldsymbol{w}, a)$ have been simplified to linear regression, we are satisfied with the asymptotic properties of $\widehat{\gamma}_{Z,n}$, $\widehat{\gamma}_{Y,n}$ (Craven and Islam 2011). To measure the accuracy of $\widehat{\mu}_n(\boldsymbol{x}, \boldsymbol{w})$, we look at the mean squared error between $G = \Phi(E)$ and $\widehat{G} = \Phi(\widehat{E})$ and between $\widehat{\mu}_n(\boldsymbol{x}, \boldsymbol{w})$ and $\mu(\boldsymbol{x}, \boldsymbol{w})$. The results are found in Figure 5.1. In both cases, the mean squared error decreases as n and p increases. Figure 5.2 contains the true and estimated piecewise linear spline for $\mu_n(\boldsymbol{x}, \boldsymbol{w}) = h(G)$. While, for the most part, the estimated spline is close to the true function, there is deviation in

the tails of the distribution.

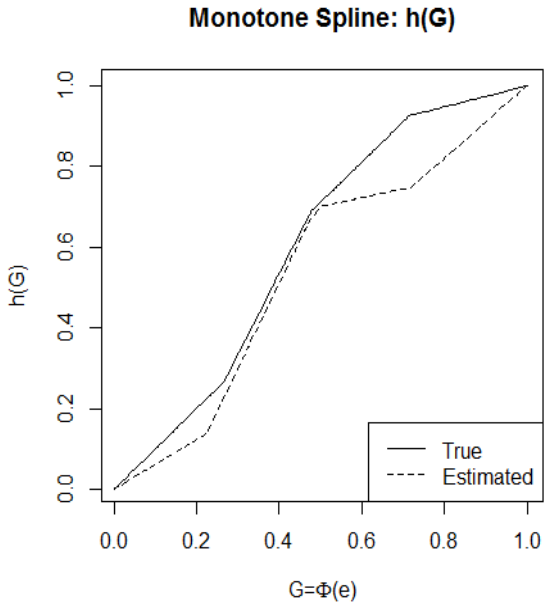
Figure 5.1: Mean squared error: $G = \Phi(E)$ and $\widehat{\mu}_n(\boldsymbol{x}, \boldsymbol{w})$



a : MSE of $\widehat{G} = \Phi(\widehat{E})$

b : MSE of $\widehat{\mu}_n(\boldsymbol{x}, \boldsymbol{w})$

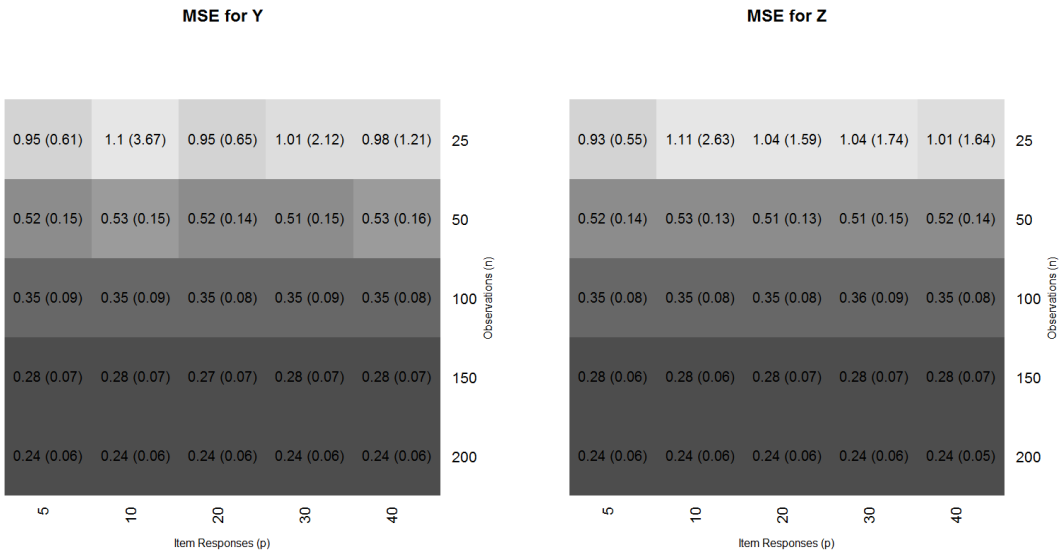
Figure 5.2: Estimated and true piecewise linear splines for $\mu_n(\boldsymbol{x}, \boldsymbol{w}) = h(G)$.



We obtained the mean squared error between the estimated value function and

the true value functions for the marginal outcomes. The results are shown in Figure 5.3 which shows the average difference between $\widehat{V}_Y(\widehat{\pi}_n) = \mathbb{E}\left[\max_a \widehat{Q}_Y(\boldsymbol{X}, \boldsymbol{W}, a)\right]$ and $V_Y(\pi_n^{\text{opt}})$ and between $\widehat{V}_Z(\widehat{\pi}_n) = \mathbb{E}\left[\max_a \widehat{Q}_Z(\boldsymbol{X}, \boldsymbol{W}, a)\right]$ and $V_Z(\pi_n^{\text{opt}})$ on the right. As expected, the quality of the approximation improves as n increases but is insensitive to changes in p. This is what we would expect because neither of the outcomes are affected by the itemized questionnaire responses.

Figure 5.3: Mean squared error: $\widehat{V}_Y(\widehat{\pi}_n)$ and $\widehat{V}_Z(\widehat{\pi}_n)$

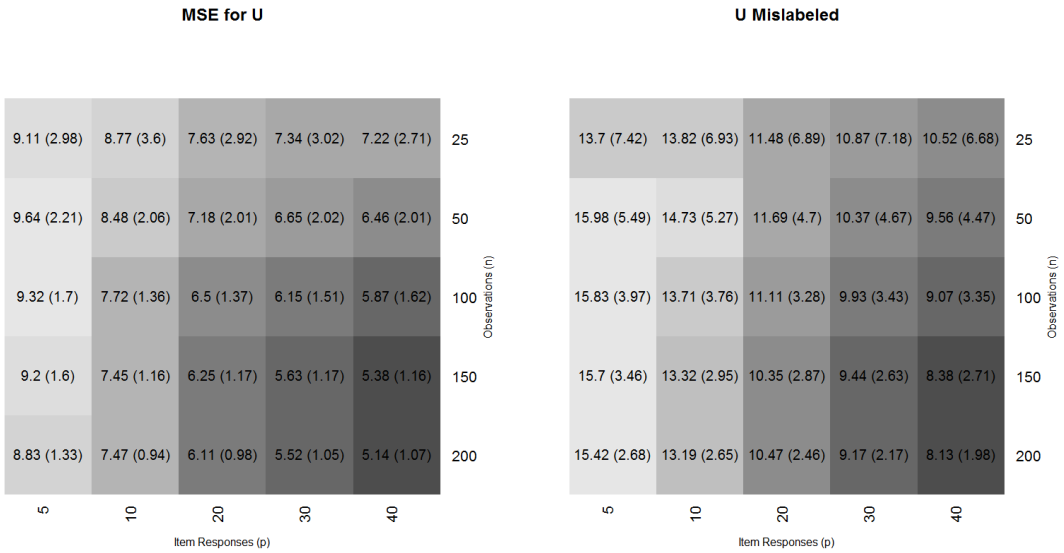


a : MSE of $\widehat{V}_Y(\widehat{\pi}_n)$

b : MSE of $\widehat{V}_Z(\widehat{\pi}_n)$

Figure 5.4 contains the mean squared error between the estimated value function $\widehat{V}_U(\widehat{\pi}_n) = \mathbb{E}\left[\max_a \widehat{Q}_U(\boldsymbol{X}, \boldsymbol{W}, a)\right]$ and the true value function of the optimal rule, $V_U(\pi^{\text{opt}})$ on the left and the percent of times the optimal treatment was mislabeled when compared to the true optimal treatment rule on the right. Except for when $n = 25$ and $p = 5$, we see that proposed estimation method performs better as sample size and items diverge. When $n = 25$ there is no discernible trend as p increases. There is a similar pattern when we fix $p = 5$ and look as n increases. These values may be too small to have desirable estimation properties.

Figure 5.4: Mean squared error: $\widehat{V}_U(\widehat{\pi}_n)$ and π^{opt}



a : MSE of $\widehat{V}_U(\widehat{\pi}_n)$

b : Average disagreement between $\widehat{\pi}_n$ and π^{opt}

5.4 Discussion

We have presented a nonparametric approach to estimating ITRs. The nonparametric assumption is employed in two separate steps of the estimation procedure to ensure reasonable flexibility within the estimator. In this instance, the benefit of this nonparametric model is increased accuracy when estimating the patient's conditional latent preference. While the logical form of the utility function remains a linear function of the outcomes, employing a nonparametric model to determine the "weights" allows a more accurate trade-off in that it accounts for the patient's satisfaction of the resulting outcomes after receiving the treatment. The method also maintains desirable asymptotic properties since it becomes more accurate as we increase the sample size and the number of items on the questionnaire.

While this method is rich in its development and represents a clear advancement in

the field of precision medicine, there are many potential extensions and enhancements. A reasonable next step is to expand this to 3 or more outcomes. This will impact the construction of the utility function, as well as the estimation of the latent trait. We also anticipate the added complexity of the model will make the algorithm more computationally intensive. This method could also be extended to consider a non linear spline component or another form of nonparametric estimation. Finally, it would be interesting to see the impact of loosening some restrictions on independence as it may be pragmatic to incorporate patient covariate information, and not just the itemized response information, into the spline estimation. Although this work provides a path in a promising direction, it is not without limitations. To ensure computational efficiency, only a piecewise linear spline was employed, and this broken stick model may not be the best representation of the underlying trends. The EM algorithm employed to estimated the conditional mean of the latent preference information is contingent on the parametrically estimated starting values for the preference information. There may be a better way to develop starting values that will lead to robust parameter estimates, such as jittering the starting values, but it has not been explored for our purposes. It also may be interesting to develop another way to incorporate the patient's satisfaction with the outcomes after receiving treatment aside from what we have already described.

CHAPTER 6: DISCUSSION

The goal of this research was to develop a set of estimation strategies that integrate patient preferences when balancing two competing outcomes. There is a logical fluidity and progression throughout the work presented here. The first setting is the single stage estimation of individualized treatment rules. Upon evaluation of this method and its extension to the multiple stage setting, it was clearly necessary to calibrate how satisfied the patient is with the results of their treatment assignment and incorporate that in the later stage estimation model. While the preference information provides a measure of how much the patient conceptually prefers one outcome over the other, the patient's contentment measures if the treatment resulted in desirable outcomes. Moving forward with the nonparametric approach, this contentment measure was included for the same reasons, even though the nonparametric model is only currently developed for the single stage setting. In application, it may be valuable to incorporate the contentment measure in the parametric model as well.

While there are many nonparametric estimation strategies, the one chosen for this work was a monotone spline. Splines were chosen because of their flexibility in the number of knots assigned and the ease of interpretation and implementation. There is great potential in comparing multiple nonparametric estimation procedures to determine which are the most computationally efficient as well as the most precise. Linear splines were chosen for the model presented because of the asymptotic guarantees defined and verified in the parametric model. There is more theoretical research to be

done to verify these asymptotic properties for quadratic or cubic splines. The non-parametric model was also only defined here for single stage estimation. It is ideal to develop an extension that performs this estimation for sequential decision making. It is theorized that a similar evolving preference model with Q-learning would be the best implementation, but further research is required to delve into the semantics of the estimation model. The form of the contentment measure only allows for a binary response of "content" or "not content". It may be beneficial to allow contentment to be measured on an ordinal scale or as an aggregate measure on a series of items, perhaps as analyzed through item response theory.

There is also vast interest in extending this work to include more than two outcomes, which requires a latent trait model for polytomous data. These can be integrated using Samejima's graded response model and the generalized partial credit model. Ideally, both methods would be compared to determine which performs the best. Samejima's graded response model is an extension of traditional item response theory but instead of assuming dichotomous responses, it assumes ordered polytomous categories (Samejima 1972). The graded response data consists of a set of items where the score of each item is an ordinal number ranging from 0 to m. The model measures the cumulative probability that person i responds to category k or lower to the j^{th} item. Conversely, the generalized partial credit model hinges on the assumption that the probability of a subject selecting the k^{th} category over all other categories is controlled by the traditional dichotomous latent trait model (Muraki 1992). Both of these models have presence in current literature, so a more extensive investigation would be needed to determine which provides more desirable properties. While these are two clear choices to develop this model for 3 or more outcomes, it is quite possible that there may be a better model all together.

One of the fundamental assumptions of this work is that estimation of the preference information is independent of the covariates and the estimation of the conditional marginal outcomes is independent of the itemized response data. This requires both the covariates and the itemized response data to be rich enough to contain sufficient information without the presence of the other. This requires diligent and thorough collection of the data, which may not always be possible. While it is reasonable to make these assumptions for this research, relaxing them may lead to a more generalizable model.

Finally, because of the potential practical significance of this work, it is of utmost importance to design, fund and implement trials that produce the well defined data required for this work. The ideal data structure only requires a small amount of additional information to collect and its collection is relatively simple and inexpensive. This will allow for not only refinement of the methodology, but more importantly provide operationalized treatment decisions that will only further advance precision medicine.

APPENDIX A: DETAILS FOR CHAPTER 3

A.1 Consistency Proof Details

We will first show that $\widehat{\beta}_n = (\widehat{\beta}_{n,0,1}, \widehat{\beta}_{n,1,1}, \ldots, \widehat{\beta}_{n,0,p}, \widehat{\beta}_{n,1,p})$ is consistent for $\beta = (\beta_{0,1}, \beta_{1,1}, \ldots, \beta_{0,p}, \beta_{1,p})$. Arcones (2006) and Fu, Li, and Zhao (1993) provide consistency results for the Rasch model that do not directly address the particulars of this model and do not allow for a large increase in sample size. The work presented here is intended to be a more generalized result. We need the following lemma and remark before giving the consistency proof in Theorem 1 below.

Lemma A.1.1. *For each $j : 1 \le j \le p$, let $X^n_{1,j}, \ldots, X^n_{nj}$ be $\overset{iid}{\sim}$ Bernoulli(π^n_j). Let $\widehat{\pi}^n_j = n^{-1} \sum_{i=1}^{n} X^n_{ij}$. Then,*

$$\max_{1 \le j \le p} |\widehat{\pi}^n_j - \pi^n_j| \to_p 0, \ \ as \ n \to \infty, \tag{6.1}$$

provided $p = o(e^n)$.

Proof. Let $\widehat{F}^n_j(t) = n^{-1} \sum_{i=1}^{n} 1\{X^n_{ij} \le t\} \ \ \forall \ \ t \in \mathbb{R}, 1 \le j \le p$. Since $\widehat{\pi}^n_j = 1 - \widehat{F}^n_j(\frac{1}{2})$, the result follows from Theorem 15.12 of Kosorok (2008). ■

Remark 1: Provided $\liminf_{n \to \infty} \min_{1 \le j \le p} \pi^n_j > 0$, (6.1) can be strengthened to

$$\max_{1 \le j \le p} |\frac{\widehat{\pi}^n_j}{\pi^n_j} - 1| \to_p 0, \ \text{as n} \to \infty,$$

provided $p = o(e^n)$.

Theorem A.1.2. *Assume*

(i) we have $3 \le p = o(e^n)$ *items with* β_0 *satisfying that* $\exists\ 0 < \delta < \frac{1}{2}$ *so that* $-\delta^{-1} \le \beta_{0,j,0} \le \delta^{-1}$ *and* $\delta \le \beta_{1,j,0} \le \delta^{-1}\ \ \forall\ \ 1 \le j \le p$ *and all* $n \ge 1$*;*

(ii) we observe n *iid individuals with* p *items,* $\{N_{ij}, 1 \le j \le p\}$*, where, conditional on latent trait* $e \sim N(0,1)$*,* $P(N_{ij}|e) = P_j(\beta_0, e)$*,* $1 \le j \le p$*.*

Then $\exists$ *an estimator of* β_0*,* $\widehat{\beta}_n$*, such that*

$$\max_{1 \le j \le p}(|\widehat{\beta}_{0,j,n} - \beta_{0,j,0}| \vee |\widehat{\beta}_{1,j,n} - \beta_{1,j,0}|) \to_p 0,\ \ as\ n \to \infty.$$

Proof. For each n, divide the p items into m groups $K_l, 1 \le l \le m$, where the groups contain between three and $3 < M < \infty$ item indices such that $K_l \subset \{1, \ldots, p\}$, there are no duplicated indices across the m groups, and that $\bigcup_{l=1}^{m} K_e = \{1, \ldots, p\}$ and $\sum_{l=1}^{m} k_l = p$, where $k_l = \#K_l, 1 \le l \le m$ (i.e., each index is represented one and only one time in the m sets). If p is fixed, then M can be set equal to p and $m = 1$. MLE estimation will be done for β_0 in subgroups defined by K_l. More specifically, for $1 \le l \le m$, we will estimate $\{\beta_{0,j}, j \in K_l\}$ though maximizing

$$(\beta_j, j \in K_l) \mapsto \sum_{i=1}^{n} \log\left(\int_{\mathbb{R}} \exp\left\{\sum_{j \in K_l} N_{ij}(\beta_{0,j} + \beta_{1,j}e) - \log\left(1 + e^{\beta_{0,j} + \beta_{1,j}e}\right)\right\} \Phi(e)\mathrm{d}e\right), \quad (6.2)$$

where $\beta_j = (\beta_{0,j}, \beta_{1,j})$ and ϕ is the standard normal density.

Let $U_l = \{0,1\}^{K_l}$ and for each $u \in U_l$, where $u = (u_j, j \in K_l)$, define $N_u = \#\{i : N_{ij} = u_j, j \in K_l\}$, where the $j's$ in K_l are in ascending order (for easier tracking).

For each $u \in U_l$, also define

$$P_u\beta) = \int_{\mathbb{R}} \prod_{j \in K_l} \left(\frac{e^{\beta_{0,j}+\beta_{1,j}e}}{1+e^{\beta_{0,j}+\beta_{1,j}e}} \right)^{u_j} \left(\frac{1}{1+e^{\beta_{0,j}+\beta_{1,j}e}} \right)^{1-u_j} \Phi(e)\mathrm{d}e.$$

Then, (6.2) can be rewritten as

$$(\beta_j, j \in K_l) \mapsto \sum_{u \in U_l} N_u \log P_u(\beta).$$

Take $\widehat{\beta}_n = (\widehat{\beta}_{j,n}, 1 \le j \le p)$ to be the maximizers of

$$(\beta_j, j \in K_l) \mapsto \sum_{u \in U_l} \frac{N_u}{n} \log P_u(\beta), \quad \forall \ \ 1 \le l \le m,$$

and note that by the consequences of being maximizers,

$$\sum_{u \in U_l} \frac{N_u}{n} \log P_u(\beta_0) \le \sum_{u \in U_l} \frac{N_u}{n} \log P_u(\widehat{\beta}_n) \le 0, \quad \forall 1 \le l \le m. \tag{6.3}$$

Note that the requirements on β_0 ensure that $\exists \ \ 0 < \epsilon < \frac{1}{2}$ such that $\epsilon \le P_u(\beta_0) \le 1-\epsilon \ \ \forall \ \ u \in U_l, \ \ 1 \le l \le m$, and all $n \ge 1$. Thus both

$$\sum_{u \in U_l} \log P_u(\beta_0) = Op(1) \text{ and } \sum_{u \in U_l} \log P_u(\widehat{\beta}_n) = Op(1), \tag{6.4}$$

where the $Op(1)$ is universally bounded for all $1 \le l \le m$. This follows from the trapping of $\sum_{u \in U_l} \log P_u(\widehat{\beta}_n)$ in (6.3). Lemma 3.4 and Remark 1 now imply that $N_u/n = P_u(\beta_0)(1+op(1))$, where the $op(1)$ is uniform over all $u \in U_l, 1 \le l \le m$. This combined with (6.3) and (6.4) implies that $op(1) \le \sum_{u \in U_l} P_u(\beta_0) \log \left(\frac{P_u(\widehat{\beta}_n)}{P_u(\beta_0)} \right) \le 0$, where $op(1)$ is uniform over all $u \in U_l, 1 \le l \le m$. This now implies that

$$
\begin{aligned}
op(1) &= \min_{1\le l\le m} \sum_{u\in U_l} P_u(\beta_0) \log\left(\frac{P_u(\widehat{\beta}_n)}{P_u(\beta_0)}\right) \\
&\le \max_{1\le l\le m} \sum_{u\in U_l} P_u(\beta_0) \log\left(\frac{P_u(\widehat{\beta}_n)}{P_u(\beta_0)}\right) \\
&\le 0,
\end{aligned}
$$

which implies that

$$
\max_{1\le j\le p}(|\widehat{\beta}_{0,j,n} - \beta_{0,j,0}| \vee |\widehat{\beta}_{1,j,n} - \beta_{1,j,0}|) \to_p 0, \quad \text{as n } \to \infty,
$$

by the properties of the Kullback-Leibler divergence and the identifiability of the constituent models. ■

From standard consistency results for linear regression, we know that $\widehat{E}(Y|X = x, A = a)$ is consistent for $E(Y|X = x, A = a)$ and $\widehat{E}(Z|X = x, A = a)$ is consistent for $E(Z|X = x, A = a)$. Now, let

$$
\widehat{Q}_n(x, w, a) = \widehat{E}\big(\Phi(e)|W = w)\big)\widehat{E}(Y|X = x, A = a) + \Big(1 - \widehat{E}\big(\Phi(e)|W = w)\big)\Big)\widehat{E}(Z|X = x, A = a)
$$

and

$$Q_0(x,w,a) = E\big(\Phi(e)|W=w\big)E(Y|X=x,A=a)+ \\ \Big(1-E\big(\Phi(e)|W=w\big)E(Z|X=x,A=a).$$

Then $\widehat{\pi}_n^{opt}(x,w) = \arg\max_a \widehat{Q}_n(x,w,a)$ and $\pi_0^{opt}(x,w) = \arg\max_a Q_0(x,w,a)$.

Note that $\widehat{\pi}_n^{opt}(x,w)$ is asymptotically equivalent to $\pi_0^{opt}(x,w)$ if the expectation for the value function for $\widehat{\pi}_n^{opt}(x,w)$ is asymptotically equivalent to the expectation of the value function for $\pi_0^{opt}(x,w)$ as $n,p \to \infty$. Hence, it is sufficient to show

$$E_{X,W}\big(Q_0(X,W,\widehat{\pi}_n^{opt}(X,W))\big) - E_{X,W}\big(Q_0(X,W,\pi_0^{opt}(X,W))\big) \to_p 0 \text{ as n } \to \infty, \quad (6.5)$$

provided we require $p \to \infty$ as $n \to \infty$.

Lemma A.1.3. *If* $\max_a(E_{X,W}|\widehat{Q}_n(X,W,a) - Q_0(X,W,a)|) \to_p 0$ *then (6.5) holds.*

Proof. Let

$$\begin{aligned} E_{X,W}\big(\Delta_n(X,W)\big) &\equiv E_{X,W}\big(\max_a|\widehat{Q}_n(X,W,a) - Q_0(X,W,a)|\big) \\ &\le E_{X,W}\big(\sum_a|\widehat{Q}_n(X,W,a) - Q_0(X,W,a)|\big) \\ &\le \#(A)\max_a\big(E_{X,W}|\widehat{Q}_n(X,W,a) - Q_0(X,W,a)|\big) \end{aligned}$$

where $\#(A)$ is the number of treatment choices which is assumed to be finite. Thus, $E_{X,W}\big(\Delta_n(X,W)\big) \to 0$ from the assertion in Lemma 2.

By definition of π_0^{opt} and $\widehat{\pi}_n^{opt}$, for any X,W, and since $\widehat{Q}_n(X,W,\pi_0^{opt}(X,W)) \le$

$\widehat{Q}_n(X, W, \widehat{\pi}_n^{opt}(X, W))$, we have:

$$
\begin{aligned}
0 &\le Q_0\big(X, W, \pi_0^{opt}(X, W)\big) - Q_0(X, W, \widehat{\pi}_n^{opt}(X, W)) \\
&= Q_0\big(X, W, \pi_0^{opt}(X, W)\big) - \widehat{Q}_n(X, W, \pi_0^{opt}(X, W)) + \widehat{Q}_n(X, W, \pi_0^{opt}(X, W)) \\
&\quad - \widehat{Q}_n(X, W, \widehat{\pi}_n^{opt}(X, W)) + \widehat{Q}_n(X, W, \widehat{\pi}_n^{opt}(X, W)) - Q_0(X, W, \widehat{\pi}_n^{opt}(X, W)) \\
&\le 2\Delta_n(X, W).
\end{aligned}
$$

Hence,

$$
E_{X,W}|Q_0(X, W, \widehat{\pi}_n^{opt}(X, W)) - Q_0\big(X, W, \pi_0^{opt}(X, W)\big)| \le 2E\big(\Delta_n(X, W)\big) \to_p 0, \text{ as n } \to \infty,
$$

and thus (6.5), and therefore Lemma 2, holds. ■

Lemma A.1.4. *Assume the following:*

(i) $\sup_w |\widehat{E}(\Phi(e)|W = w) - E(\Phi(e)|W = w)| \to_p 0, \quad \text{as} \quad n \to \infty;$

(ii) *(a)* $\max_a E_X|\widehat{E}(Y|X, A = a) - E(Y|X, A = a)| \to_p 0, \text{ as } n \to \infty;$

(b) $\max_a E_X|\widehat{E}(Z|X, A = a) - E(Z|X, A = a)| \to_p 0, \text{ as } n \to \infty.$

Then $\max_a E_{X,W}|\widehat{Q}_n(X, W, a) - Q_0(X, W, a)| \to_p 0, \quad \text{as} \quad n \to \infty.$

Proof. This follows directly from (i) and (ii) combined with the definitions of $\widehat{Q}_n$ and Q_0. ■

The following theorem gives us the desired consistency of both $\widehat{E}(\Phi(e)|W = w)$ and $\tilde{E}(\Phi(e))$:

Theorem A.1.5. *Assume:*

(i) $p \to \infty$ *as* $n \to \infty$*;*

(ii) $\exists\ 0 < \delta < \frac{1}{2}$ *such that* $|\beta_{0,j}| \le \delta^{-1}$ *and* $\delta \le \beta_{i,j} \le \delta^{-1}$ *for all* $1 \le j \le p$ *and all* $p \ge 1$*;*

(iii) $\max_{1\le j\le p}(|\widehat{\beta}_{n,0,j} - \beta_{0,j}| \vee |\widehat{\beta}_{n,1,j} - \beta_{1,j}|) \to_p 0$, *as* $n \to \infty$*;*

Then:

(a) $\max_w |\widehat{E}\big(\Phi(e)|W = w\big) - E\big(\Phi(e)|W = w\big)| \to_p 0,$

(b) $\max_w |E\big(\Phi(e)|W = w\big) - \Phi\big(\tilde{e}_p(\beta_0, w)\big)| \to_p 0,$

(c) $\max_w |\Phi\big(\tilde{e}_p(\widehat{\beta}_n, w)\big) - \Phi\big(\tilde{e}_p(\beta_0, w)\big)| \to_p 0,$

as $n \to \infty$*.*

Proof. The assumptions imply that $|\widehat{\beta}_{0,j,n}| \le 2/\delta$ and $\delta/2 \le \widehat{\beta}_{n,1,j} \le 2/\delta$ with probability approaching 1 as $n \to \infty$. To achieve convergence in probability, we can assume it holds hereafter for all n without loss of generality.

Let $q_p(e, \widehat{\beta}_n, w) = \sum_{j=1}^{p} \widehat{\beta}_{1,j,n} w_j e - \sum_{j=1}^{p} \log(1 + e^{\widehat{\beta}_{0,j,n} + \widehat{\beta}_{1,j,n} e})$ and note that

$$\widehat{E}(\Phi(e)|W = w) = \frac{\int_{\mathbb{R}} \Phi(e) e^{q_p(e,\widehat{\beta}_n,w)} \Phi(e) \mathrm{d}e}{\int_{\mathbb{R}} e^{q_p(e,\widehat{\beta}_n,w)} \Phi(e) \mathrm{d}e}.$$

Note also that

$$\frac{\partial}{\partial e} q_p(e, \widehat{\beta}_n, w) = \sum_{j=1}^{p} \widehat{\beta}_{1,j,n}(w_j - P_j(\widehat{\beta}_n, e))$$

and

$$(\frac{\partial}{\partial e})^2 q_p(e, \widehat{\beta}_n, w) = -\sum_{j=1}^{p} \widehat{\beta}_{1,j,n}^2 P_j(\widehat{\beta}_n, e)\big(1 - P_j(\widehat{\beta}_n, e)\big) < 0.$$

Hence, $e \mapsto q_p(e, \widehat{\beta}_n, w)$ has a unique maximum at $\tilde{e}_p(\widehat{\beta}_n, w)$. Now define

$$\tilde{q}_p(e, \widehat{\beta}_n, b) = \sum_{j=1}^{p} \widehat{\beta}_{1,j,n} \left(eb - \frac{\sum_{j=1}^{p} \log(1 + e^{\widehat{\beta}_{0,j,n} + \widehat{\beta}_{1,j,n} e})}{\sum_{j=1}^{p} \widehat{\beta}_{1,j,n}} \right),$$

and let $e'_p(\beta, b)$ be the unique value of e solving

$$\sum_{j=1}^{p} \beta_{1,j} b = \sum_{j=1}^{p} \beta_{1,j} P_j(\beta, e),$$

for any $b \in [0,1]$. Then

$$\sup_w |\widehat{E}\left(\Phi(e)|W = w\right) - \Phi\left(\tilde{e}_p(\widehat{\beta}_n, w)\right)| \leq$$
$$\sup_{b \in [0,1]} \left| \frac{\int_{\mathbb{R}} \Phi(h) e^{\tilde{q}_p(e, \widehat{\beta}_n, b) - \tilde{q}_p(e'_p(\widehat{\beta}_n, b), \widehat{\beta}_n, b)} \phi(e) \mathrm{d}e}{\int_{\mathbb{R}} e^{\tilde{q}_p(e, \widehat{\beta}_n, b) - \tilde{q}_p(e'_p(\widehat{\beta}_n, b), \widehat{\beta}_n, b)} \phi(e) \mathrm{d}e} - \Phi(e'_p(\widehat{\beta}_n, b)) \right| \equiv \sup_{b \in [0,1]} \widehat{A}_n(b).$$

Therefore, once we show that $\sup_{b \in [0,1]} \widehat{A}_n(b) \to 0$ and (c), then we are able to conclude (a).

Fix $b \in (0,1)$. Note that $\tilde{q}_p(e, \widehat{\beta}_n, b) - \tilde{q}_p(e'_p(\widehat{\beta}_n, b), \widehat{\beta}_n, b) \leq 0 \ \forall \ e$ with equality only if $e \equiv e'_p(\widehat{\beta}_n, b)$. From previous derivations,

$$\frac{-\log\left|\left(\frac{b}{1-b}\right)\right| - 2/\delta}{\delta/2} \leq e'_p(\widehat{\beta}_n, b) \quad \leq \quad \frac{\left|\log\left(\frac{b}{1-b}\right)\right| + 2/\delta}{\delta/2}; \tag{6.6}$$

and, moreover,

$$P_j(\widehat{\beta}_n, e)(1 - P_j(\widehat{\beta}_n, e)) \quad \geq \quad \frac{e^{-2/\delta - (2/\delta)|e|}}{1 + e^{-2/\delta - (2/\delta)|e|}} \left(\frac{1}{1 + e^{2/\delta + (2/\delta)|e|}} \right) \equiv e(\delta, e), \ \forall \ e \in \mathbb{R}.$$

Thus:

$$0 < e(\delta, h)\frac{\delta}{2} \quad \leq \quad \frac{\sum_{j=1}^{p} \widehat{\beta}_{1,j,n}^2 P_j(\widehat{\beta}_n, e)\big(1 - P_j(\widehat{\beta}_n, e\big)}{\sum_{j=1}^{p} \widehat{\beta}_{1,j,n}} \ \forall\ e \in \mathbb{R}.$$

Hence, for any $e \in \mathbb{R}$,

$$\frac{\tilde{q}_p(e, \tilde{\beta}_n, b) - \tilde{q}_p(e'_p(\tilde{\beta}_n, b), \tilde{\beta}_n, b)}{\sum_{j=1}^{p} \tilde{\beta}_{1,j}} \leq \frac{-\delta e(\delta, \tilde{e})}{4}(e - e'_p(\tilde{\beta}_n, b))^2,$$

where $\tilde{e}$ is on the line segment between e and $e'_p(\tilde{\beta}_n, b)$. By (6.6), $e'_p(\tilde{\beta}_n, b)$ is uniformly bounded for any fixed $b \in (0,1)$ and for all $p \geq 1$. Hence for any $e \neq e'_p(\tilde{\beta}_n, b)$, $p\left[\tilde{q}'_p(e, \tilde{\beta}_n, b) - \tilde{q}'_p(e'_p(\tilde{\beta}_n, b), \tilde{\beta}_n, b)\right] \to_p \infty$. Thus, $\forall\ \ b \in (0,1), \widehat{A}_n(b) \to_p 0$.

Letting $\widehat{B}_n(b) = \dfrac{\int_{\mathbb{R}} \Phi(e) e^{\tilde{q}_p(e, \tilde{\beta}_n, b)} \Phi(e) \mathrm{d}e}{\int_{\mathbb{R}} e^{\tilde{q}_p(e, \tilde{\beta}_n, b)} \Phi(e) \mathrm{d}e}$, we have that

$$\frac{\partial}{\partial b}\widehat{B}_n(b) = \frac{\int_{\mathbb{R}} \Phi(e)(e - t'_p(\widehat{\beta}_n, b)) e^{\tilde{q}_p(e, \tilde{\beta}_n, b)} \Phi(e) \mathrm{d}e}{\int_{\mathbb{R}} e^{\tilde{q}_p(e, \tilde{\beta}_n, b)} \Phi(e) \mathrm{d}e}, \tag{6.7}$$

where $t'_p(\beta, b) = \dfrac{\int_{\mathbb{R}} e^* e^{\tilde{q}_p(e, \beta, b)} \Phi(e) \mathrm{d}e}{\int_{\mathbb{R}} e^{\tilde{q}_p(e, \beta, b)} \Phi(e) \mathrm{d}e}$.

Hence (6.7) is the covariance between two monotonically increasing functions of e, which implies (6.7) > 0. Hence $\widehat{B}_n(b)$ is monotonically increasing in b. Similarly it is also easy to show that $\Phi(e'_p(\widehat{\beta}_n, b))$ is monotonically increasing in b.

By definition of $e'_p(\widehat{\beta}_n, b)$, we have that

$$\frac{\partial}{\partial e} e'_p(\widehat{\beta}_n, b) = \left(\frac{\sum_{j=1}^{p} \widehat{\beta}^2_{n,1,j} P_j(\widehat{\beta}_n, e)\big(1 - P_j(\widehat{\beta}_n, e)\big)}{\sum_{j-1}^{p} \widehat{\beta}_{n,1,j}} \right)^{-1},$$

and thus for any $0 < \rho < \frac{1}{2}$, both

$$\limsup_{n\to\infty} \sup_{b\in[\rho,1-\rho]} \frac{\partial}{\partial b} e'_p(\widehat{\beta}_n, b) < \infty$$

and

$$\limsup_{n\to\infty} \sup_{b\in[\rho,1-\rho]} \frac{\partial}{\partial b} e'_p(\beta_0, b) < \infty,$$

where both $\frac{\partial}{\partial b} e'_p(\tilde{\beta}_n, b)$ and $\frac{\partial}{\partial b} e'_p(\beta_0, b)$ are also > 0 for all $b \in [0,1]$ and all $p \geq 1$. Therefore, $\forall\ \rho > 0$, $\exists\ 0 < b_1 < \ldots < b_k < 1$ such that

$$0 \leq \Phi(e'_p(\beta_0, b_{j+1})) - \Phi(e'_p(\beta_0, b_j)) \leq \rho \tag{6.8}$$

$\forall\ 0 \leq j \leq k$, where $b_0 = 0$ and $b_{k+1} = 1$. Previous arguments show that $\Phi(e'_p(\widehat{\beta}_n, b)) - \Phi(e'_p(\beta_0, b)) \to_p 0$, as $n \to \infty\ \forall\ b \in (0,1)$. This, combined with (6.8) and the monotonicity of $b \mapsto \Phi(e'_p(\tilde{\beta}_n, b))$, yields that

$$\sup_{b\in[0,1]} |\Phi(e'_p(\widehat{\beta}_n, b)) - \Phi(e'_p(\beta_0, b))| \to_p 0, \text{ as n } \to \infty. \tag{6.9}$$

This result, combined with the fact that $\widehat{B}_n(b)$ in monotone in b and contained in $[0,1]$, as well as the fact that $\widehat{A}_n(b) \to_p 0\ \forall\ b \in (0,1)$, implies that $\sup_{b\in[0,1]} \widehat{A}_n(b) \to_p$, as $n \to \infty$.

This now implies that

$$\max_w |\widehat{E}(\Phi(e)|W = w) - \Phi(\tilde{e}_p(\widehat{\beta}_n, w))| \to_p 0. \tag{6.10}$$

The previously presented smoothness results now imply that for all $\{a_n\}, \{b_n\} \in [0,1]$ such that $|a_n - b_n| \to_p 0$, $\Phi(e'_p(\widehat{\beta}_n, a_n)) - \Phi(e'_p(\beta_0, b_n)) \to_p 0$, since (6.9) implies

$$\max_w \left| \Phi\left(\tilde{e}'_p\left(\widehat{\beta}_n, \frac{\sum_{j=1}^p \widehat{\beta}_{1,j,n} w_j}{\sum_{j=1}^p \widehat{\beta}_{1,j,n}} \right) \right) - \Phi\left(\tilde{e}'_p\left(\beta_0, \frac{\sum_{j=1}^p \widehat{\beta}_{1,j,n} w_j}{\sum_{j=1}^p \widehat{\beta}_{1,j,n}} \right) \right) \right| \to_p 0. \tag{6.11}$$

Note that previous arguments yield

$$\max_w \left| \frac{\sum_{j=1}^p \widehat{\beta}_{1,j,n} w_j}{\sum_{j=1}^p \widehat{\beta}_{1,j,n}} - \frac{\sum_{j=1}^p \beta_{1,j,0} w_j}{\sum_{j=1}^p \beta_{1,j,0}} \right| \to_p 0.$$

Now suppose that

$$\max_w \left| \Phi\left(\tilde{e}'_p\left(\beta_0, \frac{\sum_{j=1}^p \widehat{\beta}_{1,j,n} w_j}{\sum_{j=1}^p \widehat{\beta}_{1,j,n}} \right) \right) - \Phi\left(\tilde{e}'_p\left(\beta_0, \frac{\sum_{j=1}^p \beta_{1,j,0} w_j}{\sum_{j=1}^p \beta_{1,j,0}} \right) \right) \right| \not\to_p 0. \tag{6.12}$$

Then $\exists$ a sequence $\{w^n\}$ such that

$$\frac{\sum_{j=1}^p \widehat{\beta}_{1,j,n} w_j^n}{\sum_{j=1}^p \widehat{\beta}_{1,j,n}} - \frac{\sum_{j=1}^p \beta_{1,j,0} w_j^n}{\sum_{j=1}^p \beta_{1,j,0}} \not\to_p 0.$$

Since this contradicts (6.7), (6.12) must not be true. Hence, by (6.11), $\max_w |\Phi\big(\tilde{e}_p(\widehat{\beta}_n, w)\big) -$

$\Phi\big(\tilde{e}_p(\beta_0, w)\big)| \to_p 0$, and (c) follows.

Combining these results with (6.10), we now have that $\max_w |\widehat{E}(\Phi(e)|W = w) - \Phi(\tilde{e}_p(\beta_0, w))| \to_p 0$ as $n \to \infty$. All of the above arguments hold true if $\widehat{\beta}_n$ is replaced with β_0. This then implies $\max_w |E(\Phi(e)|W = w) - \Phi(\tilde{e}_p(\beta_0, w))| \to_p 0$ as $n \to \infty$. Thus, (b) and (a) follow, hence the proof is complete. ■

Remark 2: Theorem 2 demonstrates that we can approximate $\widehat{E}(\Phi(e)|W = w)$ with $\Phi(\tilde{e}_p(\widehat{\beta}_n^+, w))$ uniformly on w where $\widehat{\beta}_n^+ = (\tilde{\beta}_{0n}, (\widehat{\beta}_{1n})^+)$ since, clearly,

$$\max_{1 \le j \le p} |\widehat{\beta}_{1,j,n}^+ - \widehat{\beta}_{1,j,n}| \to_p 0.$$

A.2 Comparison of $\widehat{\mu}_{E,n}^{MH}(\boldsymbol{x}, \boldsymbol{w})$ and $\widehat{\mu}_{E,n}^{MoM}(\boldsymbol{x}, \boldsymbol{w})$ Details

We first compared the absolute difference between the Metropolis Hastings and method of moments estimator, $|\widehat{\mu}_{E,n}^{MH}(\boldsymbol{x}, \boldsymbol{w}) - \widehat{\mu}_{E,n}^{MoM}(\boldsymbol{x}, \boldsymbol{w})|$. The results are found in Figure 6.1. They show that regardless of n and p, the absolute difference is the greatest for extreme values of $\boldsymbol{W}$ (closest to 0 and 1) and is much smaller in between. Across n, the absolute difference between the two methods decreases as p increases.

Figure 6.1: Absolute difference: $\widehat{\mu}_{E,n}^{MH}(\boldsymbol{x},\boldsymbol{w})$ and $\widehat{\mu}_{E,n}^{MoM}(\boldsymbol{x},\boldsymbol{w})$

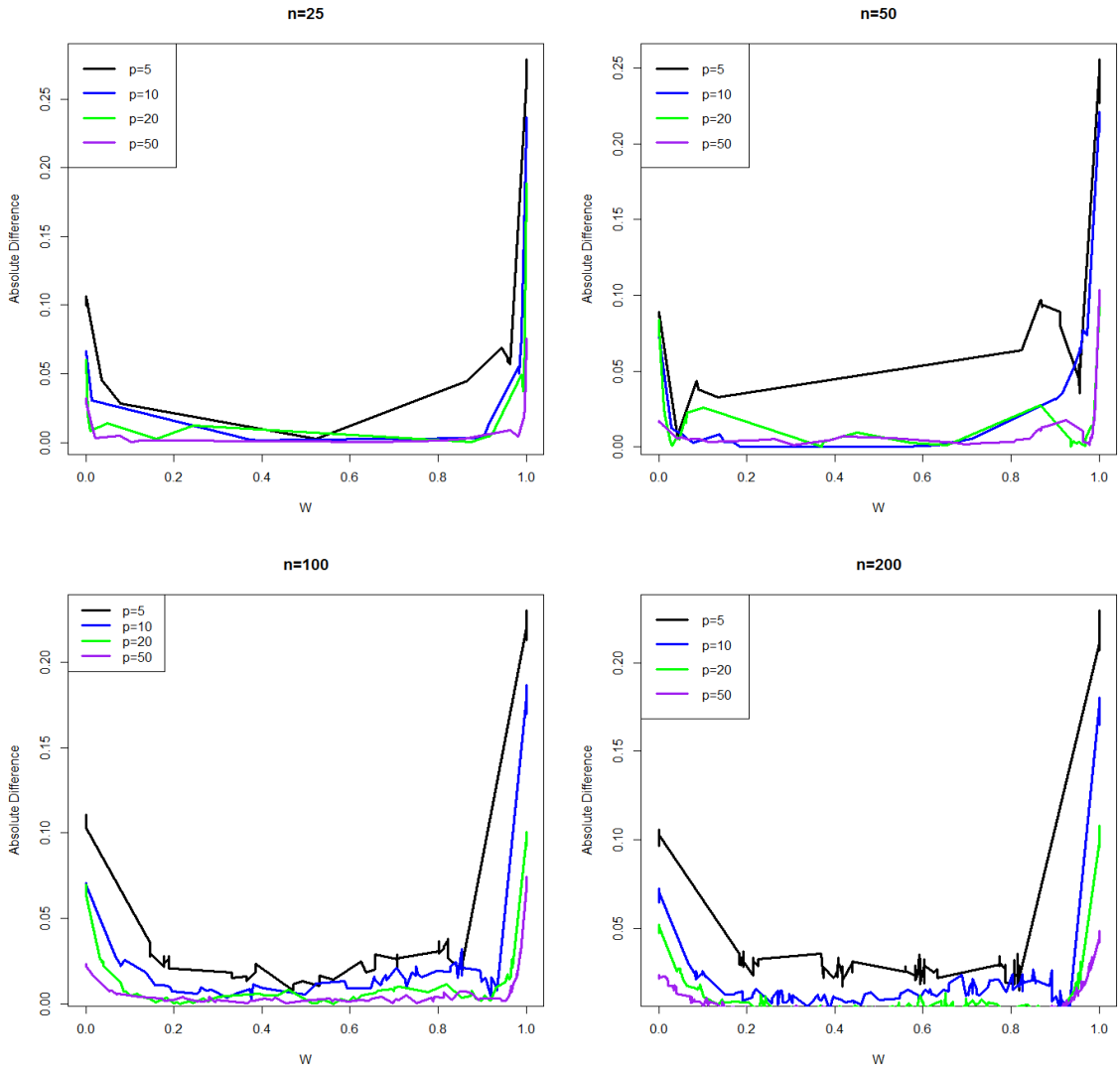
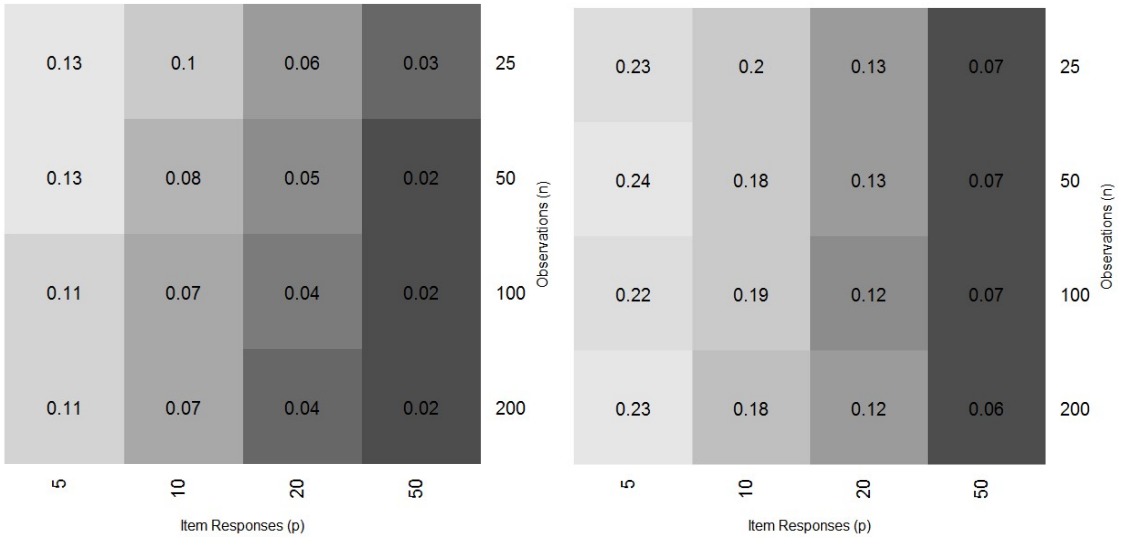


Figure 6.2 contains a heatmap of the average absolute difference on the left and a heatmap of the average maximum absolute difference between $\widehat{\mu}_{E,n}^{MH}(\boldsymbol{x},\boldsymbol{w})$ and $\widehat{\mu}_{E,n}^{MoM}(\boldsymbol{x},\boldsymbol{w})$ on the right. In both plots we see that as n and p increase, the absolute difference decreases. Collectively, these simulations provide further evidence that the methods of moments estimator performs as well as the Metropolis Hastings estimator.

Figure 6.2: Averaged absolute and mean difference: $\widehat{\mu}_{E,n}^{MH}(\boldsymbol{x},\boldsymbol{w})$ and $\widehat{\mu}_{E,n}^{MoM}(\boldsymbol{x},\boldsymbol{w})$



a : Average Absoute Difference

b : Maximum Absolute Difference

A.3 Case Study Details

PANSS was the primary assessment instrument the investigators used to assess psychopathology. The patient is rated from 1 to 7 on 30 different schizophrenic symptoms based on a face to face interview. Values range from 30-154, where larger values imply more severe schizophrenia symptoms. The difference in the PANSS score is the PANSS score at the discontinuation of phase 1 minus that PANSS score at baseline. For the sake of the proposed method, the outcome variable is this difference times -1. If the PANSS score decreases between baseline and discontinuation of phase 1 the difference will be a positive number and if it increases the difference will be negative. Hence, larger values are associated with a more favorable outcome.

The second outcome is an aggregate of indicators of side effects and adverse events,

recorded from multiple sources including systematic inquiry adverse events, serious adverse events, parameters associated with metabolic syndrome, and assessment scales for tardive dyskinesia and akathisia. Each adverse event or side effect was assigned a particular weight based on presence and severity. Adverse events were collected via a systematic inquiry where the patient was able to report adverse events such as constipation, dry mouth and hypersomnia at the discontinuation of phase 1. Although we chose to use the information from the last visit, another strategy that could be used, which is consistent with adverse event reporting standards for clinical trials would be to weight the scoring based on the most severe reporting of the symptom during the phase. Using the patient's indication of the severity of the adverse event, each is weighted as either 1/3 for mild, 2/3 for moderate, or 1 for severe. Serious adverse events were collected via a spontaneous inquiry (or whenever they arose). The serious adverse event is recorded if the event occurred within phase 1 or within 30 days of ending phase 1. The most serious event, death, was removed from this analysis because it would have been weighted so heavily as to skew the outcome. Serious adverse events include hospitalizations and even though hospitalizations for psychosis indicate a lack of efficacy, we chose to include all hospitalizations. An additional option would be to remove hospitalizations for psychosis as a serious adverse event. All serious adverse events were given a weight of 1, unless they resulted in discontinuation of the medication (hence discontinuing phase 1) then the event was weighted as a 2. Measures of metabolic effects were designated as adverse events if they exhibited a significant negative change from baseline. (It was only marked as an adverse event if the patient was not already exhibiting the negative effect at baseline). For this analysis, the last available measure was included. They will each be weighted as follows: a pulse rate above 100 was weighted as 1; a waist-hip ratio between 0.81 and 0.85 was weighted 1/2 while greater than 0.85 received a weight of 1; a blood glucose level between 100 mg/dL and 125 mg/dL was weighted 1/2 and greater

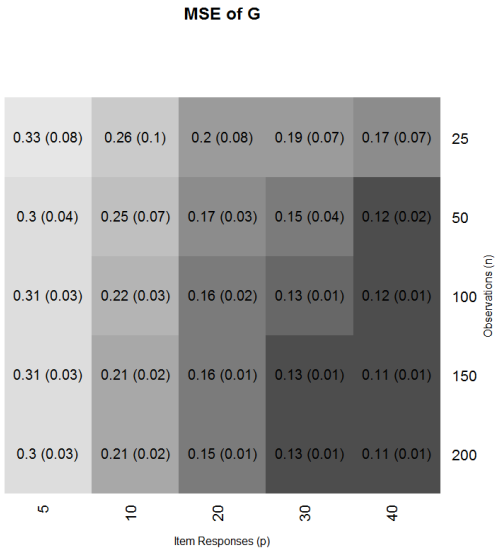
than 125 mg/dL was weighted 1; total cholesterol between 200 mg/dL and 239 mg/dL was weighted 1/2 and greater than 239 mg/dL is weighted 1; HDL cholesterol less than 40 mg/dL was weighted as 1; triglycerides between 150 mg/dL and 199 mg/dL were weighted 1/3, 200 mg/dL to 499 mg/dL was weighted 2/3 and 500 mg/dL or higher was weighted 1; patients who experienced a weight gain of 7% or higher were weighted as 1. If a patient meets the Schooler-Kane criteria for tardive dyskinesia, they were weighted as 1. If akathisia is present it is listed as mild, moderate, marked, or severe and was weighted as 1/4, 1/2, 3/4, or 1, respectively. These measurement weights were then summed up for each patient. Developing the second outcome in this manner makes larger values less desirable because larger values imply more adverse events. Since the goal is to reverse this so that larger values are less desirable, the outcome is defined as $Z = |Z^* - \max(Z^*)|$. We note that other adverse events were reported by spontaneous report and were not included in the aggregate side effect calculation because the known medical concerns for these FDA-approved medications were reasonably covered by the systematic inquiry AEs, SAEs, parameters associated with metabolic syndrome, and the assessments for movement disorders

APPENDIX B: DETAILS FOR CHAPTER 5

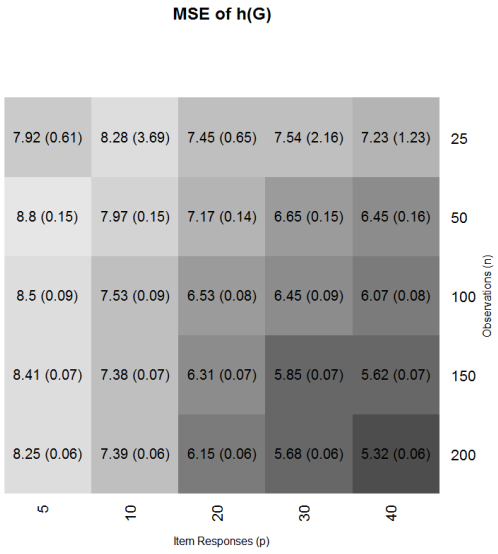
B.1 Simulation Results for 5 Knots

Below we find the simulations results plots when we include 5 knots in the monotone spline function. We see that the more complex model does not provide substantially more accurate estimates compared to the models with 3 knots.

Figure 6.3: Mean squared error: $G = \Phi(E)$ and $\widehat{\mu}_n(\boldsymbol{x}, \boldsymbol{w})$

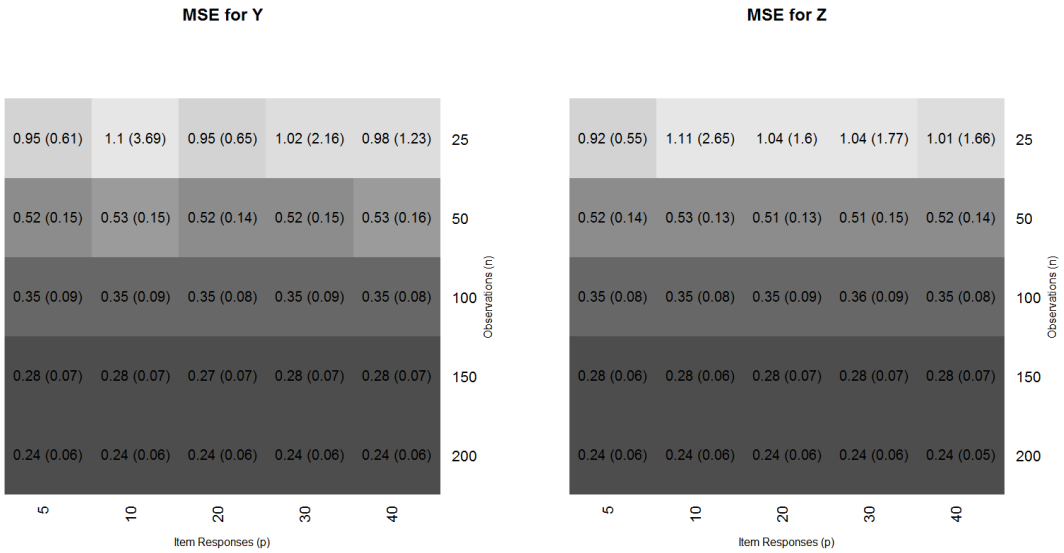


a : MSE of $\widehat{G} = \Phi(\widehat{E})$



b : MSE of $\widehat{\mu}_n(\boldsymbol{x}, \boldsymbol{w})$

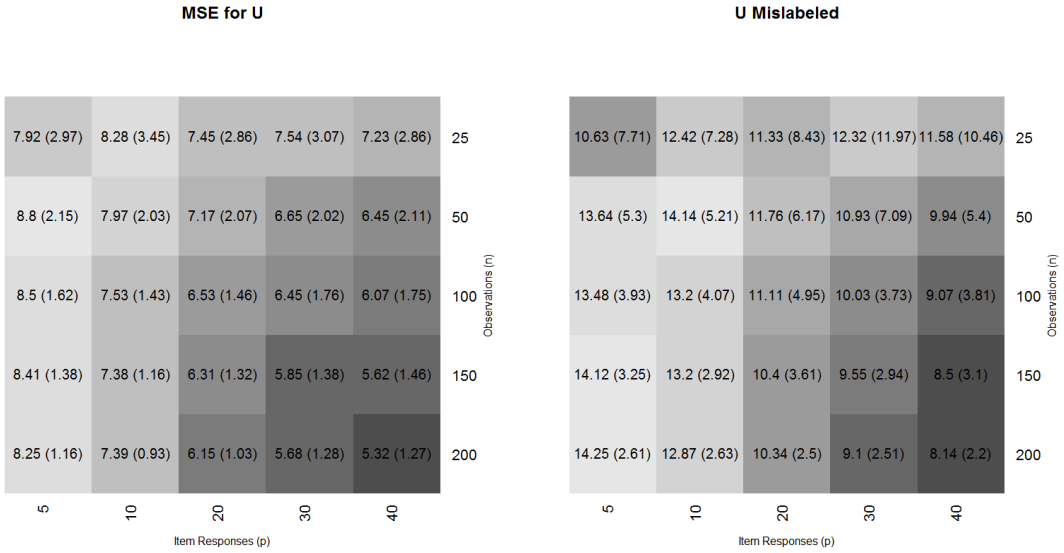
Figure 6.4: Mean squared error: $\widehat{V}_Y(\widehat{\pi}_n)$ and $\widehat{V}_Z(\widehat{\pi}_n)$



a : MSE of $\widehat{V}_Y(\widehat{\pi}_n)$

b : MSE of $\widehat{V}_Z(\widehat{\pi}_n)$

Figure 6.5: Mean squared error: $\widehat{V}_U(\widehat{\pi}_n)$ and the average percent difference: π^{opt}



a : MSE of $\widehat{V}_U(\widehat{\pi}_n)$

b : Average disagreement between $\widehat{\pi}_n$ and π^{opt}

REFERENCES

Allen, T. M. and Cullis, P. R. (2004), "Drug delivery systems: entering the mainstream," *Science*, 303, 1818–1822.

Almirall, D., Compton, S. N., Gunlicks-Stoessel, M., Duan, N., and Murphy, S. A. (2012), "Designing a Pilot Sequential Multiple Assignment Randomized Trial for Developing an Adaptive Treatment Strategy," *Statistics in Medicine*, 31, 188–192.

An, X. and Yung, Y.-F. (2014), "Item Response Theory: What It Is and How You Can Use the IRT Procedure to Apply It," in *Proceedings of the SAS Global Forum 2014 Conference*, SAS Institute, Inc.

Andreasen, N. C. and Flaum, M. (1991), "Schizophrenia: the characteristic symptoms." *Schizophrenia bulletin*, 17, 27.

Arcones, M. A. (2006), "Large deviations for M-estimators," *Annals of the Institute of Statistical Mathematics*, 58, 21–52.

Barry, M. J. and Edgman-Levitan, S. (2012), "Shared decision making—the pinnacle of patient-centered care," *New England Journal of Medicine*, 366, 780–781.

Basu, A. and Meltzer, D. (2007), "Value of information on preference heterogeneity and individualized care," *Medical Decision Making*, 27, 112–127.

Biernot, P. and Moodie, E. E. (2010), "A Comparison of Variable Selection Approaches for Dynamic Treatment Regimes," *The International Journal of Biostatistics*, 6, 1557–4679.

Braziunas, D. (2006), "Computational approaches to preference elicitation," *Department of Computer Science, University of Toronto, Tech. Rep.*

Brennan, P. F. (1998), "Improving health care by understanding patient preferences," *Journal of the American Medical Informatics Association*, 5, 257–262.

Cai, T., Tian, L., Wong, P. H., and Wei, L. (2011), "Analysis of randomized comparative clinical trial data for personalized treatment selections," *Biostatistics*, 12, 270–282.

Chakraborty, B., Laber, E. B., and Zhao, Y. (2013), "Inference for Optimal Dynamic Treatment Regimes Using an Adaptive m-Out-of-n Bootstrap Scheme," *Biometrics*, 69, 714–723.

Chakraborty, B. and Moodie, E. E. (2013), *Statistical Methods for Dynamic Treatment Regimes*, Springer.

Chen, G., Zeng, D., and Kosorok, M. R. (2016), "Personalized Dose Finding Using Outcome Weighted Learning," *Journal of the American Statistical Association.*

Collins, F. S. and Varmus, H. (2015), "A new initiative on precision medicine," *New England Journal of Medicine*, 372, 793–795.

Collins, L. M., Murphy, S. A., Nair, V. N., and Strecher, V. J. (2005), "A Strategy for Optimizing and Evaluating Behavioral Interventions," *Annals of Behavioral Medicine*, 30, 65–73.

Collins, L. M., Murphy, S. A., and Stretcher, V. (2007), "The Multiphase Optimization Strategy (MOST) and the Sequential Multiple Assignment Randomized Trial (SMART): New Methods for More Potent eHealth Interventions," *American Journal of Preventative Medicine*, 32, 112–118.

Collins, L. M., Nahum-Shani, I., and Almirall, D. (2014), "Optimization of behavioral dynamic treatment regimens based on the sequential, multiple assigment, randomization trial (SMART)," *Clinical Trials*, 11, 426–434.

Craven, B. and Islam, S. M. (2011), *Ordinary least-squares regression*, SAGE Publications.

Davis, S. M., Koch, G. G., Davis, C. E., and LaVange, L. M. (2003), "Statistical approaches to effectiveness measurement and outcome-driven re-randomizations in the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) studies," *Schizophrenia Bulletin*, 29, 73–80.

Davis, S. M., Stroup, T. S., Koch, G. G., Davis, C. E., Rosenheck, R. A., and Lieberman, J. A. (2011), "Time to all-cause treatment discontinuation as the primary outcome in the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) schizophrenia study," *Statistics in Biopharmaceutical Research*, 3, 253–265.

De Ayala, R. J. (2013), *The theory and practice of item response theory*, Guilford Publications.

Drake, R. E., Deegan, P. E., and Rapp, C. (2010), "The promise of shared decision making in mental health." .

Edwards, A. and Elwyn, G. (2009), *Shared decision-making in health care: Achieving evidence-based patient choice*, Oxford University Press.

Embretson, S. E. and Reise, S. P. (2013), *Item response theory*, Psychology Press.

Frank, R. G. and Zeckhauser, R. J. (2007), "Custom-made versus ready-to-wear treatments: Behavioral propensities in physicians' choices," *Journal of Health Economics*, 26, 1101–1127.

Fu, J., Li, G., and Zhao, D. (1993), "On large deviation expansion of distribution of maximum likelihood estimator and its application in large sample estimation," *Annals of the Institute of Statistical Mathematics*, 45, 477–498.

Ghalanos, A. and Theussl, S. (2015), *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*, r package version 1.16.

Goldberg, Y. and Kosorok, M. R. (2012), "Q-learning with Censored Data," *Annals of Statistics*, 40, 529–560.

Goldberg, Y., Song, R., and Kosorok, M. R. (2013), "Adaptive Q-learning," *Institute of Mathematical Statistics Collections*, 9, 150–162.

Hamburg, M. A. and Collins, F. S. (2010), "The path to personalized medicine," *New England Journal of Medicine*, 363, 301–304.

He, X. and Shi, P. (1998), "Monotone B-spline smoothing," *Journal of the American statistical Association*, 93, 643–650.

Henderson, R., Ansell, P., and Alshibani, D. (2010), "Regret-Regression for Optimal Dynamic Treatment Regimes," *Biometrics*, 66, 1192–1201.

Hodgkin, D., Volpe-Vartanian, J., Merrick, E. L., Horgan, C. M., Nierenberg, A. A., Frank, R. G., and Lee, S. (2012), "Customization in prescribing for bipolar disorder," *Health economics*, 21, 653–668.

Hogan, T. P., Awad, A., and Eastwood, R. (1983), "A self-report scale predictive of drug compliance in schizophrenics: reliability and discriminative validity," *Psychological medicine*, 13, 177–183.

Holland, P. W. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–960.

Huskamp, H. A., O'Malley, A. J., Horvitz-Lennon, M., Taub, A. L., Berndt, E. R., and Donohue, J. M. (2013), "How quickly do physicians adopt new drugs? The case of second-generation antipsychotics," *Psychiatric Services*.

Jain, K. (2002), "Personalized medicine." *Current opinion in molecular therapeutics*, 4, 548–558.

Jameson, J. L. and Longo, D. L. (2015), "Precision medicine—personalized, problematic, and promising," *Obstetrical & Gynecological Survey*, 70, 612–614.

Kay, S. R., Flszbein, A., and Opfer, L. A. (1987), "The positive and negative syndrome scale (PANSS) for schizophrenia." *Schizophrenia bulletin*, 13, 261.

Kidwell, K. M. (2014), "SMART designs in cancer research: Past, present, and future," .

Kosorok, M. R. (2008), *Introduction to empirical processes and semiparametric inference*, Springer Science & Business Media.

Laber, E. B., Linn, K. A., and Stefanski, L. A. (2014a), "Interactive model building for Q-learning," *Biometrika*, asu043.

Laber, E. B., Lizotte, D. J., and Ferguson, B. (2014b), "Set-valued Dynamic Treatment Regimes for Competing Outcomes," *Biometrics*, 70, 53–61.

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E., and Murphy, S. A. (2014c), "Dynamic Treatment Regimes: Technical Challenges and Applications," *Electronic Journal of Statistics*, 8, 1225–1272.

Laber, E. B. and Zhao, Y. (2015), "Tree-based methods for individualized treatment regimes," *Biometrika*, 102, 501–514.

Lavori, P. W. and Dawson, R. (2014), "Introduction to Dynamic Treatment Strategies and Sequential Multiple Assignment Randomization," *Clinical Trials*, 11, 393–399.

Li, Y. and Baron, J. (2012), *Behavioral Research Data Analysis with R*, Springer New York, chap. Item Response Theory, pp. 139–159.

Li, Z. and Murphy, S. A. (2011), "Sample size formulae for two-stage randomized trials with survival outcomes," *Biometrika*, 98, 503–518.

Lieberman, J. A., Stroup, T. S., McEvoy, J. P., Swartz, M. S., Rosenheck, R. A., Perkins, D. O., Keefe, R. S., Davis, S. M., Davis, C. E., Lebowitz, B. D., et al. (2005), "Effectiveness of antipsychotic drugs in patients with chronic schizophrenia," *New England Journal of Medicine*, 353, 1209 – 1223.

Linn, K. A., Laber, E. B., and Stefanski, L. (2016), "Estimation of dynamic treatment regimes for complex outcomes: balancing benefits and risks," in *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*, eds. Kosorok, M. R. and Moodie, E. E., New York: SIAM, chap. 15, pp. 249–260.

Lizotte, D. J., Bowling, M., and Murphy, S. A. (2012a), "Linear fitted-q iteration with multiple reward functions," *The Journal of Machine Learning Research*, 13, 3253–3295.

— (2012b), "Linear Fitted-Q Iteration with Multiple Reward Functions," *Journal of Machine Learning Research*, 13, 3253–3295.

Lizotte, D. J. and Laber, E. B. (2016), "Multi-Objective Markov Decision Processes for Data-Driven Decision Support," *Under review*, 1–25.

Lu, W., Zhang, H. H., and Zeng, D. (2013), "Variable Selection for Optimal Treatment Decision," *Statistical Methods in Medical Research*, 22, 493–504.

Moodie, E. E., Chakraborty, B., and Kramer, M. S. (2012), "Q-learning for Estimating Optimal Dynamic Treatment Rules from Observational Data," *Canadian Journal of Statistics*, 40, 629–645.

Moodie, E. E., Dean, N., and Sun, Y. R. (2014), "Q-learning: Flexible learning about useful utilities," *Statistics in Biosciences*, 6, 223–243.

Moodie, E. E. M. and Richardson, T. S. (2009), "Estimating Optimal Dynamic Regimes: Correcting Bias under the Null," *Scandinavian Journal of Statistics*, 37, 126–146.

Muraki, E. (1992), "A Generalized Partial Credit Model: Application of an EM Algorithm," 16, 159–176.

Murphy, S. A. (2005a), "An experimental design for the development of adaptive treatment strategies," .

— (2005b), "A generalization error for Q-learning," *Journal of machine learning research: JMLR*, 6, 1073.

Nahum-Shani, I. (2013), "What is a JITAI?" in *Proceedings of Workshop on Just In Time Adaptive Interventions (JITAIs)*.

Norvig, P., Relman, D. A., Goldstein, D. B., Kammen, D. M., Weinberger, D. R., Aiello, L. C., Church, G., Hennessy, J. L., Sachs, J., Burrows, A., et al. (2010), "2020 Visions," *Nature*, 463, 26–32.

Orellana, L., Rotnitzky, A., and Robins, J. M. (2010), "Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part I: main content," *The International Journal of Biostatistics*, 6.

Qian, M. and Murphy, S. A. (2011), "Performance Guarantees for Individualized Treatment Rules," *Annals of Statistics*, 39, 1180–1210.

Rasch, G. (1961), "On general laws and the meaning of measurement in psychology," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, University of California Press Berkeley, CA, vol. 4, pp. 321–333.

— (1980), *Probabilistic models for some intelligence and attainment tests*, Danish National Institute for Educational Research.

Rich, B., Moodie, E. E. M., Stephens, D. A., and Platt, R. W. (2010), "Model Checking with Residuals for g-estimation of Optimal Dynamic Treatment Regimes," *The International Journal of Biostatistics*, 6, 12–22.

Rizopoulos, D. (2006), "ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses," *Journal of Statistical Software*, 17, 1–25.

Robins, J. M. (2004), *Proceedings of the Second Seattle Symposium in Biostatistics*, Springer, chap. Optimal Structural Nested Models for Optimal Sequential Decisions, pp. 189–326.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000), "Marginal structural models and causal inference in epidemiology," *Epidemiology*, 550–560.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, 89, 846–866.

Rubin, D. B. (1978), "Bayesian inference for causal effects: The role of randomization," *The Annals of statistics*, 34–58.

Samejima, F. (1972), "A gneral model for free-response data," 18.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014), "Q-and A-learning methods for estimating optimal dynamic treatment regimes," *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29, 640.

Shortreed, S. M., Laber, E., Stroup, T. S., Pineau, J., and Murphy, S. A. (2014), "A multiple imputation strategy for sequential multiple assignment randomized trials," *Statistics in Medicine*, 33, 4202–4214.

Shortreed, S. M. and Moodie, E. E. (2012), "Estimating the optimal dynamic antipsychotic treatment regime: evidence from the sequential multiple-assignment randomized Clinical Antipsychotic Trials of Intervention and Effectiveness schizophrenia study," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61, 577–599.

Smoller, J. W. and Nierenberg, A. A. (1999), "Small numbers, big impact," *Harvard review of psychiatry*, 7, 109–113.

Sox, H., McNeil, B., Eden, J., Wheatley, B., et al. (2008), *Knowing What Works in Health Care:: A Roadmap for the Nation*, National Academies Press.

Stroup, T. S., McEvoy, J. P., Swartz, M. S., Byerly, M. J., Glick, I. D., Canive, J. M., McGee, M. F., Simpson, G. M., Stevens, M. C., and Lieberman, J. A. (2003), "The National Institute of Mental Health Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project: schizophrenia trial design and protocol development," *Schizophrenia Bulletin*, 29, 15–31.

Tao, P. D. and An, L. T. H. (1997), "CONVEX ANALYSIS APPROACH TO D. C. PROGRAMMING: THEORY, ALGORITHMS AND APPLICATIONS," *ACTA Mathematica Vietnamica*, 22, 289–355.

Taylor, J. M. G., Cheng, W., and Foster, J. C. (2015), "Reader reaction to "A robust method for estimating optimal treatment regimes" by Zhang et al.(2012)," *Biometrics*, 71, 267–273.

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014), "A simple method for estimating interactions between a treatment and a large number of covariates," *Journal of the American Statistical Association*, 109, 1517–1532.

Uebersax, J. (2000), "Latent Trait Analysis and Item Response Theory (IRT) Models," Http://john-uebersax.com/stat/lta.htm.

Villalobos, M. and Wahba, G. (1987), "Inequality-constrained multivariate smoothing splines with application to the estimation of posterior probabilities," *Journal of the American Statistical Association*, 82, 239–248.

Wang, L., Rotnitzky, A., Lin, X., Millikan, R. E., and Thall, P. F. (2012), "Evaluation of Viable Dynamic Treatment Regimes in a Sequentially Randomized Trial of Advanced Prostate Cancer," *Journal of the American Statistical Association*, 107, 493–508.

Watkins, C. J. and Dayan, P. (1992), "Q-Learning," *Machine Learning*, 8, 279–292.

Ye, Y. (1987), "Interior Algorithms for Linear, Quadratic, and Linearly Constrained Non-Linear Programming," Ph.D. thesis, Department of ESS, Stanford University.

Young, J. G., Cain, L. E., Robins, J. M., O'Reilly, E. J., and Hernan, M. A. (2011), "Comparative Effectiveness of Dynamic Treatment Regimes: An Application of the Parametric G-Formula," *Statistics in Biosciences*, 3.

Yu, C. H. (2013), "A Simple Guide to the Item Response Theory (IRT) and Rasch Modeling," .

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. (2012a), "Estimating Optimal Treatment Regimes from a Classification Perspective," *Stat*, 1, 103–114.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012b), "A Robust Method for Estimating Optimal Treatment Regimes," *Biometrics*, 68, 1010–1018.

— (2013), "Robust Estimation of Optimal Dynamic Treatment Regimes for Sequential Treatment Decisions," *Biometrika*, 100, 681–694.

Zhao, Y., Zeng, D., Laber, E. B., and Kosorok, M. R. (2014), "New Statistical Learning Methods for Estimating Opitmal Dynamic Treatment Regimes," *Journal of the American Statistical Association*.

Zhao, Y., Zeng, D., Laber, E. B., Song, R., Yuan, M., and Kosorok, M. R. (2015), "Doubly Robust Learning for Estimating Individualized Treatment with Censored Data," *Biometrika*, 102.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012), "Estimating Individualized Treatment Rules Using Outcome Weighted Learning," *Journal of the American Statistical Association*, 107, 1106–1118.

Zhao, Y., Zeng, D., Socinski, M. A., and Kosorok, M. R. (2011), "Reinforcement Learning Strategies for Clinical Trials in Nonsmall Cell Lung Cancer," *Biometrics*, 67, 1422 – 1433.